A Theoretical Model for a Dictionary of the Endangered Sherpa Language

by
Sang Yong Lee

Thesis presented in fulfilment of the requirements for the degree of
Master of Art in the Faculty of Lexicography at Stellenbosch University

Supervisor: Prof. Rufus H. Gouws

March 2017

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

March 2017

Abstract

According to Ethnologue (2015) every year an average of 6 languages are disappearing from this world, and 1,531 languages among 7,102 are classified as 'threatened' or 'shifting'. These languages are the endangered languages. This is a reason, why linguists need to help to protect the language 'species' like the protection directed at other types of categories like wild life, animals, fishes, etc. in order to preserve the species. There are a few ways to preserve the languages, which are under the danger of becoming extinct. Hinton (2001) pointed out that the most important thing to do for the endangered languages is to document the knowledge of the speakers of these languages as thoroughly as possible. Tsunoda (2005) described the theory of the holistic approach as the best way to document one language. As a method of holistic approaches dictionary compiling is a synthetic result of all grammatical analysis, and a great repository of vocabulary items. The purpose of this thesis is to present a model for a dictionary of the endangered Sherpa language as an example of how a dictionary for endangered languages can be planned and compiled.

For this purpose, attention is given to the situation of endangered languages in the world, with a focus on the importance of compiling dictionaries to revitalize the languages. The situation of the Sherpa language is explained as a sample case of some problems of documentation. Aspects of lexicography theories, i.e. dictionary typology, the theoretical approach to standard-preserving dictionaries, the general theory of Wiegand, and the function theory are discussed. Data collection is treated but it is shown that before getting involved in data collection, the language study should be done in order to ensure a clear distinction between the different dialects and variants. Two kinds of data collection are then discussed, i.e. data collection by means of a lexicographic corpus and by employing semantic domains. The structures of the envisaged Sherpa Dictionary as a model for the endangered languages are then discussed. The structural description of the Sherpa Dictionary, i.e. the outer texts (front matter and back matter), and the central list with the emphasis on the macro- and microstructure are discussed.

In this thesis, I tried to distinguish the method of dictionary compilation for the endangered languages from the method of dictionary compilation for languages that are not under the threat of extinction. This thesis will present guidelines for the envisaged Sherpa Dictionary, and will hopefully also be a model for dictionaries of other endangered languages.

Opsomming

Volgens Ethnologue (2015) verdwyn daar jaarliks gemiddeld 6 tale, terwyl 1531 van die wêreld se 7102 tale as “bedreig” of “verskuiwend” geklassifiseer is. Al hierdie tale is bedreigde tale. Daarom is dit belangrik dat taalwetenskaplikes moet help om hierdie taalspesies te beskerm – soos die beskerming van ander tipes kategorieë soos wildlewe, diere, visse, ensovoorts in ’n poging om die spesie te beskerm. Daar is verskillende maniere om tale wat met uitsterwing bedreig word te beskerm. Hilton (2001) wys daarop dat die belangrikste ding wat gedoen moet word, is om die kennis van die sprekers van daardie tale so deeglik as moontlik te dokumenteer. Tsunoda (2005) beskryf die teorie van die holistiese benadering as die beste manier om een taal te dokumenteer. As ’n metode van die holistiese benadering is woordeboeksamestelling ’n sintetiese produk van al die grammatikale analises en ’n belangrike neerslagplek van woordeskatitems. Die doel van hierdie tesis is om ’n model daar te stel vir ’n woordeboek vir die bedreigde Sherpa taal, as ’n voorbeeld van hoe ’n woordeboek vir bedreigde tale beplan en saamgestel kan word.

Vir hierdie doel is daar aandag gegee aan die situasie van bedreigde tale in die wêreld, met ’n fokus op die belang van woordeboeksamestelling om tale te laat herleef. Die situasie van Sherpa word verduidelik as ’n voorbeeld van sekere dokumentasieprobleme. Aspekte van leksikografiese teorieë, naamlik woordeboektipologie, die teoretiese benadering tot standaard-beskermende woordeboeke, die algemene teorie van Wiegand en die funksieteorie word bespreek. Dataversameling word bespreek maar dit word aangetoon dat voordat daar met dataversameling begin word daar ’n studie van die taal gedoen moet word om ’n duidelike onderskeid tussen verskillende dialekte en variante te verseker. Twee tipes dataversameling word dan bespreek, te wete dataversameling deur middel van ’n leksikografiese korpus en deur gebruik te maak van semantiese domeine. Die strukture van die beplande Sherpa woordeboek as ’n model vir bedreigde tale word bespreek. Die struktuurbeskrywing van die Sherpa woordeboek, naamlik die buitetekste (voor en agter tekste) asook die sentrale teks, met die klem op die makro- en die mikrostruktuur, word bespreek.

In hierdie tesis het ek probeer om die metode van die samestelloing van ’n woordeboek vir bedreigde tale te onderskei van die metodes vir die samestelling van woordeboeke van tale wat nie deur uitsterwing bedreig word nie. Hierdie tesis sal riglyne verskaf vir die beplande Sherpa woordeboek wat hopelik ook ’n model vir woordeboeke van ander bedreigde tale kan wees.

Acknowledgement

It has been my honor and privilege to be able to work in the field of lexicography during the past 25 years of my life. When I was young, I had never dreamed of a future as a dictionary specialist. At this point, I have come to realize that it was God who guided me in this direction.

Whenever I think of the first year of my dictionary journey, the late Prof. Ballabh Mani Dahal is the first person who comes to mind. He was my dictionary guru, and opened my eyes to the field of lexicography.

After I met the Sherpa people and got to know their language, I experienced many struggles, especially during the first five years, because of all the dialect problems. Dr. Troy Bailey was the person who taught me the methods of sociolinguistic survey. Even now, in working on this thesis, I could use valuable data from the results of that survey experience.

South Africa and the University of Stellenbosch are the best gifts in my life to enable to study the theory of lexicography even though I could well be one of the oldest students at the university. It is not easy to guess the pleasure coming from the studying after the age of 60, but I received the opportunity and have really enjoyed it. I want to give all thanks and praise to God.

Prof. Dr. Rufus H. Gouws, thankfully, opened the way for me to study at the University of Stellenbosch, and guided me practically in writing this thesis. Without him, I could not have completed this work. He is my advisor, but he has treated me as a colleague. I want to give my special thanks to him and his wife Ilse.

Eleanor J. McAlpine was an English tutor for my children, when they were young. Her faithfulness has continued for many years, and recently she volunteered to become my English editor for this thesis.

Finally, my thanks and love from the depth of my heart goes to my wife Hae Lyun, who showed her love and patience while I was busy with this study. Thanks to my son Hyun Sung, his wife Ji Hye, and my loving grandsons, Raham and Raon. Thanks to my daughter Young Lim and her husband Peter. I do not forget their prayers for me and my study.

All glory to my God!

Table of Contents

List of Figures

List of Tables

List of maps

Abbreviations

adj	adjective
adv	adverb
Ag	Agent
BNC	British National Corpus
CBSN	Central Bureau of Statistics of Nepal
COBUILD	Collins Birmingham University International Language Database
Con	connector
conj	conjugation
cons	consonant
DDP	Dictionary Development Program
de	definition
dem. pron	demonstrative pronoun
EGIDS	Expanded Graded Intergenerational Disruption Scale
Excl	Exclusive in the pronoun of the first person plural
ft	free translation
Gen	Genitive
gl	gloss
hon	honorific word
imp	imperative
Inc	Inclusive in the pronoun of the first person plural
int. pron	interrogative pronoun
is	index of semantics
lit	literal translation
LGP	language for general purposes
Loc	locative
LSP	language for special purposes
LV	light verb
lx	lexeme
mb	morpheme break

MoE	Ministry of Education
ND	Northern Dialect of Sherpa language
n	noun
nom	nominalized form
OCR	Optical Character Recognition
part	particle
pc	past conjunct
pd	past disjunct
ph	phonetic pronunciation
poss	possessive
pp	postposition
pr	present
pron	pronoun
PV	primary verb
ref	reference number
RTT	Recorded Text Test
sd	semantic domain
suf	suffix
tx	text
UNESCO	United Nations Educational, Scientific and Cultural Organization
vi	verb intransitive
vt	verb transitive
WD	Western Dialect of Sherpa language

Chapter 1. Introduction

1.0 Introductory remarks

It has been my privilege to observe part of the language shift of the Sherpa language for almost three decades, since 1988. For the Sherpas, Edmund Hillary, who in 1953 with Tenzing Norgay Sherpa was the first to climb to the peak of Mt. Everest for the first time, is more than a hero, almost a god after he committed himself to charitable work for the Sherpa people and established many local schools, health clinics, etc. in the deep Himalayan valleys. Because of all his charitable acts he should be respected, and actually he is a highly-honored person in the Sherpa community. If I am permitted to point out a small handicap in his 50 years of enormous contributions, it is the fact that he was not a linguist. Since he believed that Sherpa children should be educated at least up to primary school level to escape their poverty and to live their lives as worthy members of society, he encouraged all of his friends around the world to support Sherpa local schools in the mountains. Before democracy came to Nepal, it was felt there should be only one language, Nepali, the national language. Even though there are more than one hundred language groups in Nepal, none of these languages were accepted as standard languages, so naturally the local schools were taught only in the Nepali medium. It was not easy to find teachers to go into the mountain areas for Hillary's schools, so non-Sherpa teachers from outside had been provided compensation for their sacrifices. They were devoted teachers but did not allow the Sherpa students to speak their mother tongue at school. They were to use only Nepali. While I was there, I met many young students, who were beaten by the teachers, when they spoke their mother tongue with their peers. Unfortunately for these young folks, their mother tongue, the beautiful Sherpa language, was regarded as shameful, causing them to be beaten. Nowadays children do not speak Sherpa with their peers or even to their parents. According to the Expanded Graded Intergenerational Disruption Scale (Lewis & Simons 2010), the level of the Sherpa language is 7, which is labeled as 'Shifting'[1]. I accept, however, that it was too much to ask Hillary to understand the value of multilingual education.

[1] According to the report of UNESCO Ad Hoc Expert Group on Endangered Languages (2003: 8-9) this level is labeled as 'definitely endangered'.

The Ethnologue (Lewis et al. 2015) reports that there are 7,102 living languages in the world. The UNESCO Ad Hoc Expert Group on Endangered Languages (henceforth UNESCO) gives detailed information by quoting Bernard (1996:142), "about 97% of the world's people speak about 4% of the world's languages; and conversely, about 96% of the world's languages are spoken by about 3% of the world's people". The saddest information is that "about 90% of the languages may be replaced by dominant languages by the end of the 21st century (UNESCO 2003:2)". Based on this information, we encounter many terms such as endangered languages, language death, and extinct languages.

This thesis is about endangered languages, how to revitalize them, and what role the dictionary plays in documenting these dying languages. In the first Chapter, I will discuss 1) what are considered endangered languages and what their situation is in this world? 2) why and how languages become endangered, 3) what the role of the dictionary is in revitalizing endangered languages, and finally, the limitations of this thesis.

1.1 What are the endangered languages and what is their situation in this world?

UNESCO (2003:2) defines a language as being in danger, "when its speakers cease to use it, use it in an increasingly reduced number of communicative domains, and cease to pass it on from one generation to the next". Ken Hale, a pioneer in the study of endangered languages, recognized this problem and made this topic of endangered languages known in worldwide linguistic circles by organizing the Endangered Language Symposium at the 1991 meeting of the Linguistic Society of America. To awaken the interest of linguists, he compared endangered languages with the loss of biological diversity on this earth (Hale 1992:1-2). Krauss (1992:7-8), who also made this comparison, remarked that the really endangered or threatened mammals, for example, may be just 10%, and 5% of birds of all species.

Nevertheless, biologists are very wisely arousing international concern to protect endangered species through more than 40 international agencies and 300 national agencies as well as private organizations. Lewis & Simons (2010:11-15) report the real statistics of endangered languages

using The Expanded Graded Intergenerational Disruption Scale (EGIDS)[2]. EGIDS classified all languages in the world by using levels 0 to 10. Languages classified as level 6b and above are called endangered, and level 10 means they are completely extinct. According to *Ethnologue* (2015) the number of languages of level 6b 'threatened' and 7 'shifting' is 1,531 (22%) of the 7,102 languages. The number of languages levels 8a through 9, classified as 'dying', is 916 languages (13%). Finally, the number of languages classified as level 10 'extinct' is 367 languages, counted only from 1950. This means that every year an average of 6 languages are disappearing from this world.[3] It is significant to see that a linguist like Peter Ladefoged openly expressed his negative attitude to supporting endangered languages (Ladefoged 1992:809-811)[4]. But the problem here is that, in comparison with biologists, the majority of linguists are so unassertive that they remain quiet regarding the truth of this tragic language endangerment.

1.2 Why and how languages become endangered?

Thomason (2015:11) explains the reason for language endangerment as being language contact. No matter what the reason is for the language contact, this contact causes the language speaker to become bilingual or multilingual, which is the first step towards language shift. Then, willingly or unwillingly, the language speakers stop speaking their own language, and the language shift moves toward endangerment. She says: "This is especially obvious when the social, political, and economic relations between the two language communities in contact are markedly assymmetrical". She gives six causes that make languages endangered (Thomason 2015:19-35). The first cause is conquest. If the conqueror's language replaces the language of the conquered, the language of the latter will be in danger or die. The second is economic pressures. This economic pressure can, and does, apply not only to conquered peoples but also to voluntary immigrants and to contact situations, in which one group comes to dominate the other(s). The third is melting pots. This is a dominant cultural ideology in the United States. To many Americans, it seems obvious that merging into a single homogeneous culture means shifting to

[2] For the details, see https://www.ethnologue.com/about/language-status
[3] For the details, see https://www.ethnologue.com/endangered-languages
[4] Ladefoged (1992: 810-811) expressed his uncomfortable feelings toward the endangered languages in his paper, "But it is not for me to assess the virtues of programs for language preservation versus those of competitive programs for tuberculosis eradication, which may also need government funds".

using English and giving up their languages. The fourth is the language of politics. Although the recent promotion of language rights on the international scene has begun to discourage discriminatory laws, it is all too easy to find countries all over the world that attempt to suppress minority languages. The fifth is attitude. What people think about their language its value, its usefulness, its importance to their culture—can play a decisive role in the language's fate. The sixth, which is a very good observation to me personally, is loss of linguistic diversity via standardization[5]. When a particular dialect is chosen to form the literary standard, other local dialects will be in danger.

1.3 What is the role of the dictionary in revitalizing endangered languages?

Although there have been criticisms in the past of linguists and scholars who worked for the minority or endangered languages but misused their data and neglected the people's culture and dignity (Tsunoda 2005:216-217), it is still very important to document endangered languages. Hinton (2001:413) pointed out, "Perhaps the most important thing to do when a language is down to a few speakers is to document the knowledge of those speakers as thoroughly as possible." This is undoubtedly important not only for languages which are extinct or nearly so, but also for languages which are generally in danger.

UNESCO (2003:16) introduced nine major evaluative factors of language vitality. The ninth factor is to evaluate the urgency for documentation. If languages have comprehensive grammars, dictionaries, and extensive texts, these are 'superlative' (Grade 5 on the EGIDS scale). But, to the degree that the quality of these materials is lower and the number and length of the documents are less, the grade becomes lower, down as low as 'inadequate' (Grade 1), which means extinct. So, documentation of endangered languages is very urgent. Otherwise they will become even more endangered and even disappear entirely.

[5] For me this is an important consideration. To revitalize minority languages, the standardization of a language is very important, but we did not think deeply enough to realize that those dialects which were excluded from the standardization process would be in danger because of the standardization. This is an ironic result for those linguists who want to revitalize endangered languages through standardization. I think this is one of the most important considerations in compiling the Sherpa Dictionary not to lose dialects, which were not selected as part of the standard dialect.

What kinds of content should be documented? Tsunoda (2005:231-233) described the theory of the holistic approach, which requires the linguist to look at the language in terms of its various aspects, such as phonology, morphology, syntax, discourse, semantics, vocabulary, etc. If we consider the limits of budget and man-power, we cannot research all linguistic areas. However, since languages are so diverse, if we focus only on a certain part of a language, for example, phonology or syntax, we could lose the general understanding of the language, and furthermore we could distort it. It is also important to be holistic in considering different genres of text collections, such as narrative, descriptive, procedural, expository, hortatory, drama, and song. These holistic genres should cover most of their historical heritage, culture, religion, and philosophy. In regard to this holistic approach, the production of a dictionary is very crucial. Dictionary-making is a synthetic result of all grammatical analysis, and also provides a great repository of vocabulary items, taken from text collections of various genres, so a dictionary is a good tool for outsiders to learn the language, and for insiders to document their language. This is particularly important when native speakers need to find a certain word or when they are confronted with the difficulty of remembering words they haven't heard for many years, but which they used to hear from their parents or grandparents (Thomason 2015:142).

There are many possible ways to document endangered languages, but this thesis focuses on documentation through dictionary-making. When I studied the Sherpa language, I found that some of the grammar papers done by previous scholars were written without having first done a dialect study. My first job was to clearly distinguish the dialects and then try to establish the orthographic system. Then, using this orthography, lots of texts in different genres were collected. Having done this, I thought it was the right time to begin to build a Sherpa dictionary. While I was making the Sherpa dictionary, my goal was not just to do that one dictionary, but to make a dictionary based on a valid theory, which could be a model for dictionaries for other endangered languages. So, the theme of this thesis is a theoretical model for the best approach to making the most effective kind of dictionary for preserving an endangered language.

1.4 Limitations of this thesis

To attain this goal, there will have to be some limitations to this thesis. The situations of the endangered languages are too varied to create a simple system for a theoretical model for a dictionary that will be adequate for all situations. So, this thesis will be limited to the languages

covered by levels 6 to 7 on the EGIDS, which include 6a (Vigorous), 6b (Threatened), and 7 (Shifting). And it is also appropriate to cover in this thesis grade 2 (Fragmentary) on UNESCO's urgency list, under which UNESCO includes some grammatical sketches, word-lists, and texts useful for limited linguistic research.

The language data, which were used in this thesis, were mostly provided by two Sherpa people, Mingmar Sherpa and Tashi Sherpa, who speak the Southern dialect of Sherpa. If there is any mistake in the data, the responsibility is completely mine.
Finally, the Sherpa Grammar is still not completed. There could be some changes in this thesis, after the grammar has been completed.

1.5 Chapter summary

In this Chapter, I discussed the definition of endangered language and the prevailing situation. Lexicographers should be aware that, even while discussing the definition and situation of endangered languages, every year an average of six languages are disappearing from this world. This is the reality of endangered languages and is a significant motivation for this thesis. How should these endangered languages be revitalized? Documentation is very important for revitalization of a language. I conclude that, for a holistic approach to this documentation, the compilation of dictionaries is vital, because one of the roles of dictionaries is to revitalize languages under the threat of extinction.

Chapter 2. The Sherpa language and the problems faced in its documentation

2.0 Introduction

Since this thesis is on creating a theoretical model for dictionaries for endangered languages, it would be good to use a specific endangered language as an example. The Sherpa language, on which the writer has been working for 27 years, has most of the typical characteristics and problems encountered in endangered languages. In 2.1, a general introduction to the Sherpa language will be given, and 2.2 will cover the problems faced in the documentation.

2.1 The Sherpa language

2.1.1 Geography of the Sherpa area

Nepal extends from the Terai in the south to the high Himalayan Mountains in the north. The southern part, the Terai, is lowland belt adjoining India, and the northern area borders Tibet's high plateau. The Himalayan range extends from east to west along Nepal's northern border like a belt between Nepal and Tibet. There are many high peaks in this area, including the world's highest mountain, Mt. Everest, which is located on the northeastern border of Nepal. This world-famous mountain is located in the Sagarmatha Zone of Nepal. To the west of the Sagarmatha Zone is the Janakpur Zone. Most of the Sherpa people inhabit these two zones, where almost all the land is between 1,000 and 4,000 meters in elevation.

Most Sherpas live in the Solu-Khumbu District in the far north of the Sagarmatha Zone. This area is divided into three regions listed from north to south, Khumbu, Pharak, and Solu. The Bhote Kosi and the Dudh Kosi rivers meet south of Namche Bazaar. Khumbu is located north of this junction. From this junction, Pharak extends south on both sides of the north-to-south Dudh Kosi valley, as far as the junction of the Dudh Kosi and Deku Khola rivers. From here, Solu extends from east to west as far as Likhu Khola and to Okhaldhunga to the south.
In the northern Janakpur Zone to the west, there are two districts, Dolakha and Ramechhap. In the northern part of these two districts, there are many Sherpas living together with other ethnic groups. The Sherpas are the largest ethnic group in the Rolwaling area, which borders Tibet.

Sherpas also live in other areas, including Sankuwasabha, Taplejun, and Ilam, scattered in eastern Nepal, but the Sherpas are a minority in these areas. Some Sherpas also live in northern India around Darjeeling and in the state of Sikkim.

2.1.2 The History of the Sherpa people

As Sherpas do not have any written record of their history, it is very hard to trace their footprints. Most Sherpas believe they originally migrated from the Kham area of Tibet. It is also not clear whether they moved to the present location at one time and by one route or if it was a gradual movement following various routes. Oppitz (1973:122) mentioned that the Sherpas left their homeland, Kham of the eastern Tibetan province named Salmo Gang, because of the religious tensions with the Mongols. He reconstructed the clan history of the Sherpas, and guessed that the first year of migration was about 1533.

The name Sherpa itself gives some insight into their migration. 'sher' is the adjectival form of 'shar', which means 'east', and 'pa' is 'people'. So they may have been given this name by the Rai people who inhabited this area at that time. However, the name 'Sherpa' is used not only in Nepal, but is also used for a group of people in the Sichuan District of eastern Tibet. They call themselves Amdo Sherpa (Nagano 1980). In this case, 'Sherpa' could just refer to an 'easterner' without reference to a migration.

Regardless of the place of origin before their immigration, there are various ideas about their origin among the Sherpas. Most of the Khumbu people believe that Khumbu (the Northern part of the Sherpa area) is the site of the first Sherpa settlement, because they came to Nepal through Nangpa La[6], and arrived in Khumbu first. It is difficult to verify whether this point of entry is correct, but most Sherpas agree that further migration started from Khumbu and moved southward to Solu, and to the Ramechhap and Dolakha areas to the west. Some much smaller groups of people moved toward Sankuwasabha, Taplejun, and Ilam and even across the border to Sikkim and Darjeeling in India.

[6] Nangpa La is a high mountain pass crossing the Himalayas on the Nepal-Tibet border, of which the elevation is 19,050 ft. (https://en.wikipedia.org/wiki/Nangpa_La)

2.1.3 The Sherpa People

Even though the Sherpas are Nepali by nationality, their religious practices and culture are much closer to those of Tibet than Nepal. They are very proud of their culture and their identity as Tibetan Buddhists. Each house has its own high pole with prayer flags, and each village has both a chorten[7] and a Buddhist temple. Manings, walls of religious phrases inscribed in stone, are readily found along the paths. As their natural environment is high in altitude, mountainous, not very good for agriculture, and they have not had enough land to cultivate, from the very early days, they have been involved in trade. Namche Bazaar is one of the famous market places between the highland and lowland of Nepal. Wool and salt came from Tibet and these were traded for grains and daily necessities. This continued until the relatively recent closure of the border between Nepal and Tibet in 1954.

It was during Sir Edmund Hillary's expedition to Mt. Everest that the Sherpas became exposed to the outside world and became known around the world as trekking guides. Nowadays the Sherpas are a basic necessity for expeditions in Nepal's high mountains, and the Solu-Khumbu area has become a famous trekking route for world tourists. Because of this, many Sherpas changed their occupations to the tourist business and now run hotels, lodges, restaurants, and trekking offices. But in areas, other than the well-known tourist regions, traditional agriculture is still practiced.

The Central Bureau of Statistics of Nepal reported in 2001 that the population of the Sherpas was 154,622, which was 0.68% of the total population of Nepal.[8]

2.1.4 The Sherpa Language

The people of Nepal speak more than 120 languages[9]. Most of these languages are still under-developed. Nepal's languages are categorized into two language families, Indo-European and

[7] a monument to a distinguished Buddhist, especially/usually a lama (http://www.dictionary.com)
[8] In 2011, the Central Bureau of Statistics of Nepal reported their decennial statistics. The Sherpa population was shown as 112,946, a decrease of 27% from the previous report in 2001. Most Sherpa people do not believe this report, but insist their population is more than 200,000. This is the reason the report of CBSN of 2001 has been used/cited here.

Tibeto-Burman. Hale (1970) introduced three classifications of the Sherpa language (Ethnologue code: SCR), based on the work of Sten Konow, Robert Shafer and Voegelin. These three classifications are not the same, but they all show that the Sherpa language belongs to the central branch of the Tibetan family, which is close to Lhasa Tibetan.

The language situation is generally shifting at the moment. The older generation still uses their mother tongue, but the society, as a whole, has less intention to use it with their children. The sustainability of the Sherpa language is very much in doubt. It has already been stated in the Introduction that the EGIDS level is 7, which is labeled as 'shifting'.

2.2 Problems found in the process of the development of the Sherpa language

2.2.1 The history of the standardization of the Sherpa language

Various linguists have been working on the analysis of this language for many years. First, Hale (1970:4) categorized the Sherpa language as the Central Tibetan Subgroup of the Tibetan family under the Sino-Tibetan phylum. Gordon (1969) wrote a Sherpa Phonemic Summary, and he and Schöttelndreyer (1970) analyzed the Sherpa language as having a two-tone system with rising and falling tones[10]. Later, however, Watters (1999:54-77) pointed out that the Sherpa language has only two registers, high and low, without rising and falling. As a result of his sociolinguistic survey, Lee (2003:81-95) reported that there are three dialects of the Sherpa language. Regarding grammar, Kelly (2003:244-452) wrote a general grammar of the Sherpa language, and Greninger (2006) studied Sherpa discourse.

Sherpa linguists, with the encouragement of foreign scholars, have also shown interest in the development of their language. Ang P. Sherpa compiled the Sherpa-Nepali-English dictionary in 1999, and Gelu Sherpa wrote a paper on Sherpa Orthography in Devanagari in 2001 and one on

---

[9] The number of individual languages listed for Nepal is 122. Of these, 120 are alive and 2 are extinct. Of the living languages, 7 are institutional, 18 are developing, 32 are vigorous, 55 are in trouble, and 8 are dying (Ethnologue: http://www.ethnologue.com/country/NP).

[10] The two-tone system, with rising and falling tone, actually represents a four-tone system, with high-rising, high-falling, low-rising, and low-falling tones.

Subject-verb Agreement in Sherpa and the English language in 2005.[11] Lhakpa N. Sherpa wrote a book called *Through a Sherpa Window*, which is a glossary type of dictionary with explanations of the Sherpa culture. In 2009 Lhakpa N. Sherpa, Nicolas Tournadre, Gyurme Chodrak and Guillaume Oisel published the *Sherpa-English and English-Sherpa Dictionary with Literary Tibetan and Nepali equivalents*. I know of two more Sherpa dictionary projects that are in progress, one under the auspices of Rinpoche[12] of Tengboche Temple and the other, Ngawang W. Sherpa's dictionary, under the auspices of Tribhuvan University.

As I view the whole process of Sherpa language development in regard to comprehensive planning, there are two basic problems which we need to resolve for the standardization of the Sherpa language. The first is the dialect issue, and the second, the script issue.

Author	Topic of paper	The dialect based on	Script
Kent Gordon	Sherpa Phonemic Summary	Southern dialect	N/A
B. Schöttelndreyer	Sherpa Segmental Synopsis	Southern dialect	N/A
S. E. Watters	Tonal Contrasts in Sherpa	Southern dialect	N/A
Barbara Kelly	A Grammar and Glossary of the Sherpa Language	Northern dialect	N/A
David Greninger	Sherpa Discourse	Southern dialect	N/A
Ang P. Sherpa	Sherpa-Nepali-English Dictionary	Southern dialect	Roman
Gelu Sherpa	Sherpa Orthography in Devanagari	N/A	Devanagari
Lhakpa N. Sherpa	Through a Sherpa Window	Southern dialect	Tibetan
Nicolas Tournadre et al.	Sherpa-English & English-Sherpa	Northern dialect	Tibetan
Tengboche Temple	Sherpa Dictionary	Northern dialect	Tibetan
Ngawang W. Sherpa	Sherpa Dictionary	Western dialect	Tibetan
Ngawang W. Sherpa	Primary School Text 1-3	Western dialect	Tibetan

Table 1. Dialects and scripts on which authors based their work

From Table 1, it is very clear that a comprehensive plan has not been used in developing the Sherpa language, but the development of this language has been based on the resource person's

[11] As Sherpa is the name/one of the names of the people group, most of their family names are Sherpa.
[12] Rinpoche is considered to be a reincarnated monk, and is the most respected lama in Sherpa society.

native dialect, and whether or not the authors understood Tibetan. In the rest of this chapter, the problems related to dialect, script, and orthography will be covered in more detail.

2.2.2 The dialect issue

Lee (2003:81-95) reported in his sociolinguistic survey that there are three dialects of the Sherpa language, the Western, Southern, and Northern dialects[13]. Among these three dialects, it is very hard to judge which one is closer to the original Sherpa language, or which one is more prestigious. The main problem in Sherpa language development is that each dialect group holds strongly to their conviction that their dialect has to be chosen as the standard dialect.

The people who speak the Northern dialect have insisted that historically their ancestors came down from Tibet 600 years ago and crossed the Himalayas through the pass at Nangpa La. They believe that the first settlement was in the present area where the Northern dialect is spoken. Furthermore, they claimed that people moved from the North to the South and then from the South to the West.[14] On the basis of their view of the historical background, the speakers of the Northern dialect strongly stress that their dialect is the original and pure Sherpa language.
On the other hand, the people who speak the Southern dialect have different reasons for considering their dialect to be the standard dialect. In the modern history of the Sherpa people, the Southern area has been the center of their religion and culture. Politically, the district office is located in the Southern area. As the Southern area is lower in altitude than the North, more of the Sherpa people live in the South.

Regarding the speakers of the Western dialect, they have not been considered as being central either religiously or culturally. However, as far as population is concerned, the West has the highest population of the three areas. This is their reason for considering the Western dialect to be the standard dialect.

[13] The Sherpa spoken in some scattered areas such as Ilam, Taplejung, and Rolwaling, etc. has not been included in these three dialects.
[14] Between the North and the West there are high peaks, so they could not cross directly from the North to the West.

2.2.3 The orthography and script issue

The Sherpa language does not have an official orthography system yet, only the various systems developed by different scholars. In 1970, Schöttelndreyer (1970) described his Sherpa orthography, based on the Devanagari script. When Ngawang W. Sherpa edited the primary school textbooks I-III, he introduced Sherpa orthography in the Tibetan script. Then, Gelu Sherpa submitted his *Sherpa Language Standard Orthography in Devanagari* to the Sherpa Culture and Language Preservation and Promotion Committee in 2009.

To finalize the Sherpa orthography, two steps need to be considered. The first is to finalize the phonemic orthography. With regard to the Sherpa orthography, two groups exist, and their phonemic understandings of the Sherpa language are not the same. One group's understanding is based on the Sherpa Phonemic Summary (Gordon 1969), which was modified by Schöttelndreyer as *A Devanagari spelling system for the Sherpa Language* (Schoettelndreyer 1970). The other group consists of the Sherpa linguists, Ngawang W. Sherpa and Gelu Sherpa. The typical disagreements between these two groups are in regard to vowel length and the understanding of the open-fronted-unrounded vowel.

As stated in 1.3 regarding the holistic approach, a natural text collection representing different genres is very crucial. When the collected texts are analyzed, if the writing system is not first standardized, it will take a long time to bring together all the different spellings for the same word. In my personal experience at the Dictionary Development Program (DDP)[15] workshop in 2005, during the two-week workshop 6,500 Sherpa words were collected by DDP. The words were collected by six Sherpas, who spoke the same dialect, but the workshop took place without the benefit of a standard Sherpa orthography. Therefore, it took several months to reduce the words collected by up to 5,500 by removing words that were listed more than once because of different spellings. In the process, I was faced with a very difficult question: Who has the authority to select the standard spelling from among several for a specific word and what should be the general orthographic rules?

[15] DDP was developed by Ron Moe (2001). I was trained to use his program, and the workshop in 2005 was run according to his instructions.

The second step to consider is the issue of script. There are two options for the script to be used, Devanagari script and Tibetan script. Devanagari is the national script of Nepal; whereas the Tibetan script is the one used in Tibet, which politically belongs to China. Since the Sherpa people are very strongly Tibetan Buddhist, they are tightly bound up with the Tibetan religion and culture. Furthermore, they believe that the Tibetan script, in which all Tibetan Buddhist scriptures were written, is holy just as Latin was considered in the Middle Ages. If they lose the Tibetan script, they believe that they will lose their Tibetan Buddhist identity.

However, the reality presents quite different picture. First, the literacy rate in the Devanagari script for Sherpa speakers is 68.5%. The rate in the mountainous region is just a little less than that, 63.7% (Central Bureau of Statistics of Nepal 2011). Since CBSN has no official statistics for Tibetan script, it is difficult to guess the literacy rate in the Tibetan script. Most of those literate in Tibetan are monks or nuns, who were trained to read and write the Tibetan scriptures usually from childhood. The Tibetan Buddhist temples are the only place where one can learn this script. In these days, there is a tendency for temples to open schools to teach their children the Tibetan religion and culture in the Tibetan language, particularly those destined to be monks or nuns[16]. It is a mere guess on my part that less than 10% of the whole Sherpa population can read and write the Tibetan language in the Tibetan script. Even if I increased this figure to 20%, still the use of the Tibetan script would be relatively low.

Second, we have to think about the question as to whether using the Tibetan script would help to improve the quality of education or would cause it to deteriorate. The general situation of education in Nepal is quite backward, even though there has been a lot of progress since the coming of democracy in 1991. Furthermore, the quality and availability of education in the mountain areas is much lower than in the hills and the Terai. The general rate of graduation from primary school for the whole Nepal was 73.6% according to the Flash Report of the Ministry of Education in Nepal (2011-2012:31). Table 2 shows the dropout rate for each grade.

[16] In the Flash Report of Ministry of Education of Nepal (2011-2012:16), the number of religious schools in the mountain areas is 78.

	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Average for grades 1 to 5
Promotion rate	70.8	87.4	89	88.8	88.4	83.1
Repetition rate	21.3	8.3	7.4	7.5	5.4	11.5
Dropout rate	7.9	4.3	3.6	3.7	6.2	5.4
Cohort Graduation Rates	-	-	-	-	-	73.6

Table 2: Internal efficiency at the primary level (MoE 2011-2012:31)

Weinberg (2009:2), who observed firsthand the real situation of education using the Tibetan script in Sherpa villages, reported, "The textbooks, according to two different Sherpa language teachers, are also too hard for the students and focus too much on grammar but fail to teach Sherpa grammar effectively."

Here, "grammar" simply means how to read and write the Tibetan script. She also said, having observed the attitude of teachers who teach the students, "One Nepali teacher felt that the classes made it harder for children to learn English and Nepali, and one Sherpa teacher said that parents just wanted their children to learn English and Nepali (Weinberg 2009:3)." Just imagine those children in primary school learning three different scripts: Roman, Devanagari, and Tibetan!

The Government of Nepal allows the primary schools to use the mother tongue of the minority language groups in order to explain the subject matter more clearly. This is the policy of using a transitional language. According to the Flash Report of 2011-2012, 33 local languages were being used as the medium of instruction. The Sherpa language was one of them, and 156 classes were involved in this system (MoE 2011-2012:58). Under this policy, if they use Devanagari script for the Sherpa language, it will be a good transition to Nepali. Therefore, they start to learn the school subjects in their mother tongue using the Devanagari script. Then, with already having skill in Devanagari, they can study other higher subjects in Nepali without having to spend time to learn another script.

Here, we have been able to observe the conflict between identity versus reality. To protect the Tibetan Buddhist identity, it would be good to keep the Tibetan script for the Sherpa language, but the reality is that most of the people are illiterate in this holy script, and young students have difficulty in learning the script.

2.3 Chapter summary

In this Chapter, I discussed the Sherpa people, their living conditions, history, and their language. After the migration from Tibet, the Sherpa people moved down to the southern part of the Himalayas, and scattered to the east and west along the Himalayan range. Geographically they live at an altitude between 1,000 and 4,000 meters, and are surrounded by many mountains, higher than 6,000 meters, and by deep river valleys. The Sherpa people have been separated by these natural barriers for more than six hundred years, and this has resulted in language changes and the formation of different language variants in each of the Sherpa communities. The survey found that there are three main dialects with many variants in each dialect. Furthermore, the political situation of Nepal had earlier focused on a policy of a single national language, Nepali. This was the situation until democracy came to Nepal. The Sherpa language was easily neglected by the country of Nepal, and even by the Sherpa people themselves.

This is the real situation of endangered languages in the world. There could be differences, but most endangered languages are in a similar situation, and are under the threat of extinction. This is the reason why I wanted this thesis to become a theoretical model for the compilation of dictionaries of endangered languages for their revitalization.

Chapter 3. General lexicographic theory

3.0 Introduction

Gouws and Prinsloo (2005:1) explain that the field of lexicography today has two components, i.e. a theoretical component and a practical component. Historically the practical dictionary was started much earlier than the advent of lexicography theory. Even though a number of people have influenced the development of lexicography theory, three scholars and their contributions will be introduced here.

The first one is Ladislav Zgusta, who published the *Manual of Lexicography* [17] in 1971, which is an extremely important work and the first major publication to establish theoretical lexicography (Gouws and Prinsloo 2005:1-2). He started his book with a focus on linguistic matters such as lexical meaning, but extended his topic to include variations such as formal variation of words and variation in language, as he wanted to reflect the real usage of language in the dictionary. Then, he introduced his typology, which will be shown in more detail in 3.1, and explained his theoretical approaches with reference to monolingual and bilingual dictionaries. He should be recognized as being on the frontier of establishing a theoretical approach to the field of lexicography, but with a special interest in endangered languages. He incorporated the culture of the respective linguistic community and the variations of the language into the formal area of lexicography (Zgusta 1971:1-2, Gouws and Prinsloo 2005:1-2).

The second is Herbert Ernst Wiegand. Gouws and Prinsloo (2005:4) emphasized Wiegand's contributions to the history of lexicography, insomuch that they made a subsection named 'The Wiegand era'. Wiegand's most important contribution is in regard to lexicographic structures. Gouws and Prinsloo (2005:5) said, "Since Wiegand (1984), numerous of his [Wiegand's] publications have dealt with wide-ranging issues regarding the structure of dictionaries." Wiegand defined the role of lexicography as meta-lexicography saying, "We must bear in mind that writing on lexicography is part of meta-lexicography and that the theory of lexicography is

[17] In 1960, UNESCO offered a contract to the International Council for Philosophy and Humanistic Sciences to inquire into the situations in the domain of lexicography. After a long process, Zgusta finally prepared this book in 1971 (Zgusta 1971: 9).

not part of lexicography." He included four components under meta-lexicography, i.e. the history of lexicography, the general theory of lexicography, research on dictionary use, and criticism of dictionaries. Details of the structure of meta-lexicography will be described in 3.3.

The third is Sven Tarp. In the history of the theory of lexicography, he is the person who focused on the functions of dictionaries (Gouws and Prinsloo 2005:7). The notion of functions had already been mentioned in the earlier theories, but during more recent years more of the attention on the function of lexicography was based upon the users of dictionaries, i.e. the situation of the users and the needs and problems of users. Gouws and Prinsloo (2005:8) explained Tarp's function theory as follows (The details of his theory will be given in 3.4.):

> According to Tarp (2002:70)[18] a lexicographic function represents the assistance that a dictionary provides to a particular type of user to cover the needs of that user in a specific user situation. Bergenholtz & Tarp (2002)[19] distinguish between knowledge- and communication-oriented functions.

Compiling a dictionary for an endangered language is quite different from doing so for major or even minor languages that are not endangered. In general, the situation of most of the endangered languages continues to deteriorate. Even the basic preparation for dictionary production cannot be easily accomplished. In this case, before the actual compilation of the dictionary can be started, a detailed study of general lexicographic theory is crucial, because the proper understanding of the theories will be fundamental to making a quality dictionary based on both good lexicographic principles and the real situation of the designated language.

As this paper seeks guidance for the formulation of an envisaged theoretical model, in this chapter, the classical approach to lexicographic theory, especially in (a) the typology as seen in Zgusta (1971) and other authors, (b) the theoretical approach to standard-preserving dictionaries (Zgusta 1989), (c) the general lexicographic theory of Wiegand (1984), as well as (d) the theory of lexicographic functions (Tarp 2008), will be shown as the basis of the methodological approach. In considering these different aspects, it will be possible to find the theoretical guidelines needed for the model of a dictionary for an endangered language.

---

[18] Tarp, S. 2002. Translation dictionaries and bilingual dictionaries – two different concepts. In: *Journal of Translation Studies 7*: 59-84.

[19] Bergenholtz & Tarp (2002). Die moderne lexikographische Funktionslehre. Diskussionsbeitrag zu neuen und alten Paradigmen, die Woerterbucher als Gebrauchsgegenstaende verstehen. In: *Lexikographica* 18: 253-263.

3.1 Typological nature of the dictionary

Gouws and Prinsloo (2005:45) said, "An important issue in any lexicographic process is the decision regarding the typological nature of the dictionary to be compiled." A typology of dictionaries is a study of how to classify dictionaries according to the differences and similarities between the different dictionary types (Tarp 2008:113). This typology will be a map for users as to what dictionary to consult when confronted with lexical problems (Swanepoel 2003:44). Zgusta (1971:223) emphasized the importance of considering the different types in the preparation period by saying, "It is necessary to decide to what type the prepared dictionary should belong." Tarp (2008:101) explains the type in regard to accessibility: "The truly unique thing about dictionaries is not various types of data, but the way in which this data is made accessible, so users can quickly and easily find the exact data."

3.1.1 Different criteria of the dictionary typology

Landau (1984:7) is right to point out the difficulties in classifying all dictionaries according to the many criteria, as there is no standard, agreed-upon taxonomy for dictionaries. In this section, a number of major criteria will be shown, but it is not possible or necessary to introduce all the types of dictionary; only the classical types and some types related to endangered languages will be discussed.

3.1.1.1 Zgusta's typology

Zgusta (1971:198), first of all, differentiated encyclopedic dictionaries from linguistic ones, explaining the difference between these two types as follows:

> The latter (linguistic dictionaries) are primarily concerned with language, i.e. with the lexical units of language and all their linguistic properties;… In contradiction to this, the encyclopedic dictionaries (…) are primarily concerned with the denotata of the lexical units (words): They give information about the extra-linguistic world, physical or non-physical, and they are only arranged in the order of the words (lexical units) by which the segments of this extra-linguistic world are referred to when spoken about (Zgusta 1971:198).

The contrast appears to be quite clear. Landau (1984:7) also clarifies the difference between encyclopedic and linguistic dictionaries, "Dictionaries are about words, encyclopedias are about things." But it is still hard to understand what the clear criterion of this distinction is here. Tarp (2008:113) expressed a similar uncertainty by saying, "Although we may intuitively sense a difference between these two concepts, the difference between them has never been clear…."

However, Zgusta's subcategories (1971:199-221) under linguistic dictionaries is adequate to be able to recognize the classic types of dictionaries. He divided linguistic dictionaries into six categories as shown below:

1) the diachronic vs. synchronic dictionaries by the criterion of history
2) the historical vs. etymological dictionaries under diachronic dictionaries
3) the general vs. restricted (or special) dictionaries under synchronic dictionaries
4) the standard-descriptive vs. overall-descriptive dictionaries under general dictionaries
5) the monolingual vs. bilingual vs. multilingual dictionaries by the numbers of languages
6) the pedagogical or reverse dictionaries by the purpose

Swanepoel (2003:46) kindly produced Zgusta's six categories in the chart below:

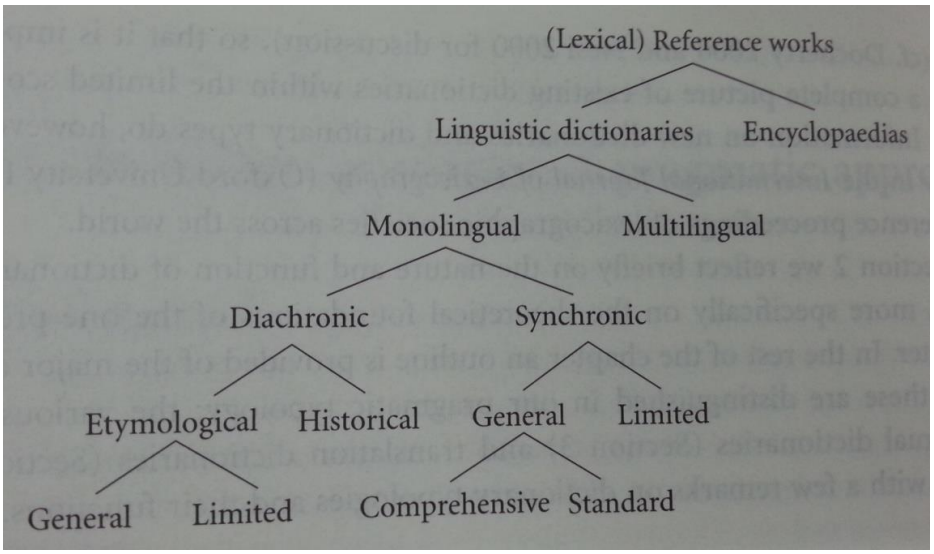


Figure 1. A dictionary typology (Zgusta 1971)

3.1.1.2 Al-Kasimi's typology

Al-Kasimi (1977:17-31) proposed three criteria for his 'new typology', (a) source, (b) scope, and (c) purpose. Even though his typology is limited to the bilingual dictionary, the first four of his seven contrasts based on the purpose are valuable for endangered languages (Al-Kasimi 1977:20, Newell 1995:11-18).

1) dictionaries for the speakers of the source language vs. speakers of the target language
2) dictionaries of the literary language vs. of the spoken language
3) dictionaries for production vs. for comprehension
4) dictionaries for the human user vs. machine translation

3.1.1.3 Tarp's typology

Tarp (2008:119-120) presented a function-oriented typology, which shows the relationship between dictionaries and their functions. Mainly he focused on the function of the learner's dictionaries. For the users wanting to learn a certain language there will be two situations such as communicative and cognitive situations. The communicative situation primarily concerns production and reception in the foreign language, and translation between the source language and the target language. The cognitive situation means that the function is to gain knowledge of the foreign language's vocabulary, grammar, or culture. On the basis of communicative and cognitive situations, he describes three types of dictionaries:

1) dictionary with communicative functions
2) dictionary with cognitive functions
3) dictionary with communicative and cognitive functions

3.1.2 Dictionary types suitable for the endangered languages

Most dictionary types were classified by considering the dictionaries of well-established languages such as English, German, and French, whose dictionaries have been developed for a couple of centuries. But most of the endangered languages do not have grammatical descriptions or many corpora from text collections, so the standard language or dialect has not yet been decided upon. Zgusta (1971:223) described this situation as the 'neuralgic problems' with which the lexicographer will be confronted. Therefore, dictionary types suitable for these endangered languages need to be considered. In this section, a few dictionary types will be introduced. These types, except 3.1.2.5, were already mentioned in 3.1.1 but are now reorganized in the light of endangered languages.

3.1.2.1 Dictionary types by different ranges or levels

Zgusta (1971:220) quotes Malkiel's classification in his footnote, as being range, perspective, and presentation. One of the subcategories under range is the density of entries. Newell (1995:9) expressed this density of entries as levels (or stages) of lexical description. These range from simple to more complex lexical description, and the size of the entries from small to big. For the endangered languages, the situation of the language development should be considered, and then decide what will be the best dictionary for the language. Newell (1995:9) classified four types with their meanings:

1) Word list: Commonly a word list results from beginning with a word list in one language and eliciting the corresponding words or phrases in an object language.

2) Glossary: A glossary is typically an alphabetically arranged list of the vocabulary used in specific texts, such as language learning books or published texts, usually appearing at the end of the material.

3) Vocabulary: This refers to a small dictionary of two to four thousand citation forms including a few sense distinctions and some illustrations, as well as possibly ethnographic descriptions of some lemmata.

4) Dictionary: A dictionary should be a relatively extensive treatment of representative lemmas of a language from four thousand citation forms onward, including ethnographic information, synonyms, antonyms, and other lexical relationships.

3.1.2.2 Dictionaries for literary vs. spoken language

Zgusta (1971:222) mentioned two basic decisions to be made at the very beginning of producing a dictionary. One of the two is to find out the varieties of language, especially the difference between the literary and the spoken form of the standard language, which is in focus. Al-Kasimi (1977:23) also explained the difference between speech and writing by saying, “Speech is the fundamental form of language activity, and writing is just a representation of speech.” After examining the existing dictionaries, he concluded that they are based on writing rather than on speech, partly because of the comparative ease of collecting written materials (Al-Kasimi 1977:23). This means that the spoken variety of language is more complicated to document than the written variety, and the documentation of the spoken variety is also not well organized. All documentations of the spoken varieties are strictly to be labeled according to the place where they were elicited, or the hometown of the person, and from whom each of the varieties were documented. Then these data should be organized by a dialect map, and be analyzed according to rules. Application using some of the rules will be as follows.

If the speech data are clearly different from the base dialect according to a certain rule, it will be relatively easy. The standard form can include the variations. I, personally, had included the speech differences in the written-oriented dictionary of the Nepali-Korean dictionary. This dictionary includes the variations in pronunciation; the spoken form first, followed by the written one as in the following examples:

(1) PAHIRANU [pairanu/pahiranu] (Lee 1999:390)[20]
(2) SAMAYA [samae/samaya] (Lee 1999:579)[21]

If the speech sources are a mixture of different dialects or regional speech forms, the more prestigious dialect, which is determined by research such as a sociolinguistic survey, has to be chosen from among the different dialects. Then the prestigious dialect can be used as a standard. If there are any grammatical rules concerning the variation between the standard and the dialects, these rules can be introduced in the front matter, so that the variations do not need to be represented by a lemma. For example, from the Sherpa language, in Table 3 the Solu and Khumbu dialect word for 'sand' has a grammatical rule that, in place of the voiceless plosive, the Khumbu dialect has aspiration. So, pema[1] in Solu dialect becomes p^hema[1] in Khumbu.[22] As these rules were already explained in the front matter, p^hema[1] did not need to be a separate lemma. However, in the online-dictionary, which is different from a paper-dictionary and has much more freedom in terms of size, it will be good to include these variations as reference articles. In this case, these reference articles will have only an article of cross-reference to show what the standard form is. If the case is not the difference in pronunciation, but a lexically different word, of course, even in the paper-dictionary it will be included as a reference article. For example, pepsok[1] 'sand' in the Western dialect, and rhilkoŋgoŋ[2] 'spider' in Khumbu.

Gloss	West	Solu	Khumbu
SAND	pepsok[1]	pema [1]	p^hema[1]
SPIDER	bəldzjaŋ[1]	bəldzjaŋ[1]	rhilkoŋgoŋ[2]

Table 3. Sample variations in Sherpa dialects (Lee 2003:85)

If the literary form and the spoken form of the language are completely different from each other, such as in the case of the Tibetan and Arabic languages, the size of lexically different words will be much bigger. According to the degree of difference between the two language varieties, the decision needs to be made whether one dictionary can cover both the literary and the spoken forms or whether two separate dictionaries should be produced one for the literary language and one for the spoken language.

[20] In modern Nepali /h/ becomes zero sound [ɸ] between vowels.
[21] In modern Nepali, /-ya/, when occurring word-finally, becomes [e].
[22] The superscripts indicate tones; 1 (low tone), 2 (high tone).

3.1.2.3 Dictionaries: Standard-descriptive vs. overall-descriptive

Zgusta (1971:211) describes the standard-descriptive dictionaries of the standard national languages, which do not include dialectical or regional words or variations of the language but describe only what is generally regular or normal. On the other hand, overall-descriptive dictionaries try to help the user to understand all texts and communications he is likely to read or hear. At the end of his explanations, he described the possibility of combining these two types in a single publication.

> The usual procedure is that the standard-descriptive part of the resulting dictionary is treated as a box within another box. The dictionary which is published is then basically of the overall-descriptive type, but all obsolete, regional etc. items are labeled as such by a sign or a label. In this way what is not labeled can be considered "normal" in the sense of a standard-descriptive dictionary (Zgusta 1971:213).

I agree with him in combining the two types, i.e. having the standard-descriptive and overall-descriptive functions in one dictionary. The Sherpa Dictionary will label the non-standard words with such terms as variation and dialect.

3.1.2.4 Dictionaries: Monolingual vs. multilingual dictionaries

The criterion for these types is the number of languages represented in the dictionary. In a monolingual dictionary, only one language is represented, which means that the object language and meta-language are the same. In a multilingual dictionary, one or more meta-languages are represented (Zgusta 1971:213, Newell 1995:9-10). Zgusta (1971:213) differentiated these two as follows:

> The usual aim of a bilingual dictionary is to help in translating from one language into another, or in producing texts in a language other than the user's native one, or both. The usual situation is that the more descriptive tasks are reserved to the monolingual dictionaries….

The Sherpa Dictionary will be a multilingual one, of which the meta-languages are Nepali and English, so that the Sherpa language will be understood by both the local Nepali people and the international reader.

3.1.2.5 Dictionaries: Based on a corpus vs. based on semantic domains

The criterion to distinguish between these types of dictionaries is the method of material collection, whether the lemmata come from the corpus of a text collection or from semantic domains. The strong point of the corpus-based dictionary is that the lemmata are from the real

language in daily use. Newell explained that a three-million-word corpus is needed for a six-thousand-word dictionary (Newell 1995:22). For endangered languages, it is not easy to obtain a corpus of one million or more words. The weak point of a corpus-based dictionary is the limitation of its lemma candidates. If the corpus is small, it is hard to find those lexical items that are not frequently used in everyday conversation.

The strong point of the semantic-domain-based dictionary is that the collection of lemmata is well balanced over most of the semantic areas. And, if the language group has well-educated people, within a workshop of a couple weeks, more than ten thousand lemmata can be gathered.[23] On the other hand, the weak point is that it is not so easy to find words that are in daily use. Also, if the people who participated in the elicitation process are not speakers of the same dialect, many variations could be added to the dictionary, which will take a long time to sort out and eliminate.

For a quality dictionary, my suggestion is to use both methods and to take advantage of the strong points of each. The detailed discussion of this will be given in Chapter Four.

3.2 Theoretical approach to standard-preserving dictionaries

3.2.1 The standard-descriptive dictionaries of Zgusta (1971)

In the process of the development of any language which has more than one dialect, one dialect becomes predominant over the other dialects (Zgusta 1971:170). In this case, there are two possibilities for what happens to the other dialects 1) they will stop being used and die out, or 2) they will be used as variations of the standard national language. If the other dialects survive and are used together with the standard national language as variations, they could be categorized in three possible ways (Zgusta 1971:120):

1) The literary language (standard national language) vs. the cultivated spoken language
2) The standard national language vs. the colloquial language
3) The standard national language vs. folk speech

[23] From my personal observation of the Tharu Word Collection Workshop on March 2005 in Nepal, instructed by Ronald Moe's Dictionary Development Program.

Here, if we focus on the standard national language only, it will be the same as the standard-descriptive dictionary of Zgusta (1971:210) as mentioned in 3.1.2.3.

3.2.2 Zgusta's role of dictionaries that influence the standard

Zgusta (2006:186-197) further developed his theory on the preservation of the standard languages in this later book. He classified dictionaries according to how the dictionaries influence the standard. His four types are as follows:

1) dictionaries that aim at creating a written standard: standard-creating dictionaries
2) dictionaries that try to render the standard more modern: modernizing dictionaries
3) dictionaries that try not only to stop any change in the standard, but even to reverse change, to reintroduce obsolete forms and meanings: antiquating (or archaizing) dictionaries
4) dictionaries that try to describe the existing standard, thereby clarifying it: standard-descriptive dictionaries

3.2.3 Standard dictionaries for endangered languages

Even though in the standard dictionaries a number of items of high frequency usage from non-standard varieties, e.g. slang, or special fields, may be included by clearly marking such as stylistic or chronological (cf. Gouws and Prinsloo 2005:50), in general, they will not include any dialects or variants (Zgusta 1971:211). For endangered languages, we do not compile a dictionary with standard lemmata, as actually there is no standard, but with unstable words, which are on the way to becoming standard. The compilation of a dictionary for endangered languages is like creating a repository of all possible lemmata that can be collected, which has their sources recorded such as the region, the dialect name, and the time of collection. Later on, with the development of the sociolinguistic survey, if the more dominant or prestigious dialect is chosen, all lemmata will be sorted according to the different dialects. In this case, the compilation of dictionaries for endangered languages has as its goal the creation of standard dictionaries. This is similar to Zgusta's standard-creating dictionaries (Zgusta 2006:187), but it is different in theory. In Zgusta's standard-creating dictionaries, 'standard' means standardizing of a foreign word (a cultural language) for the translation of a certain language (a new language) in written form, for example, the creation of the Old Church Slavonic written standard language for the translation of the Bible (Zgusta 2006:187). In this case, standardization means to create a standard word for the new foreign terms. In this paper, we are talking about the standardization from different variants of one word.

Zgusta also showed an interesting case about varieties that are raised to standards (Zgusta 2006:188). After British English had been introduced into America, Australia, New Zealand, etc., the English spoken in those countries became variants of the original British English. These variants created the need for monolingual dictionaries, for example, the Australian English Dictionary and the Dictionary of Newfoundland English (Zgusta 2006:188). In terms of standardizing among the variants, this is similar to the process of standardizing endangered languages, but it is quite different because these variants are based on British English, which already had a strong standardization.

3.3 The general theory of the lexicography of Wiegand

Most of the monumental achievements of H. E. Wiegand were written in German, except for a couple of papers published in English. Therefore, this chapter will mostly depend on one of his publications, *On the structure and contents of a general theory of lexicography* (Wiegand 1984), and on a few introductory books written in English by other people. Gouws and Prinsloo (2005:5), without hesitation, called this age the 'Wiegand era' and stated the characteristics of the Wiegand era as follows:

> The Wiegand era has been characterized by the identification of the different components of dictionary articles and by a meticulous description of their specific structure and function (Gouws and Prinsloo 2005:5).

Through forming specific structures for a dictionary, Wiegand has built up the general theory of lexicography, which covers the formal features of dictionary-making, so that this theory can be applied in lexicographic practice (Gouws and Prinsloo 2005:5).

Furthermore, he focused not only on building up the general theory of lexicography, but also on a variety of dictionary structures such as the data distribution structure, the frame structure, the macro- and microstructure, the addressing structure and the access structure. In this regard, Gouws and Prinsloo (2005:5) described Wiegand's achievement in discussing the structure of dictionaries as follows:

> In his prolific portfolio of publications Wiegand has focused dictionary research not only on the contents of dictionaries and dictionary articles but also on the structure of dictionaries. Since Wiegand (1983), numerous of his publications have dealt with wide-ranging issues regarding the structure of dictionaries (Gouws and Prinsloo 2005:5).

3.3.1 What is lexicography?

First of all, Wiegand clarified the meaning of lexicography in relationship with other subjects, such as meta-lexicography, applied linguistics, and lexicology, i.e.

> (1) Lexicography was never a science, it is not a science, and it will probably not become a science. [….] We must bear in mind that writing on lexicography is part of meta-lexicography and that the theory of lexicography is not part of lexicography.
>
> (2) Lexicography is not a branch of so-called applied linguistics. [….] lexicography is, at all events, more than the application of linguistic theories and methods or the utilization of linguistic and philosophical findings. [….]
>
> (3) Lexicography is not a branch of lexicology, and lexicography is by no means theoretically determined by lexicology alone. [….]
>
> (4) Lexicographical activities result in reference works which can be classified according to different types. All types of works made with the aim of providing not only, but above all, information on linguistic expressions should be classified as linguistic lexicography. They would include at least the following types: dictionaries of language, glossaries, concordances and word indexes [….] (Wiegand 1984:13-14)

After this clarification, he concluded about what linguistic lexicography (which he later referred to simply as lexicography) is by saying, “Linguistic lexicography is scientific practice aimed at producing reference works on language, in particular dictionaries of language.” (1984:14)
Then he introduced a general theory of lexicography and its components as follows:

> (1) The lexicographical activities. These can be classified into three fields of activity:
>
> (a) The first field includes all the activities leading to the drawing up of a dictionary plan.
>
> (b) The second field of activity includes all the activities involved in establishing a dictionary base and in processing this base in a lexicographical file.
>
> (c) The third field of activity includes all the activities concerned directly with the writing of dictionary texts and thus with the writing of the dictionary.
>
> (2) The results of the lexicographical activities in the three fields, namely: the dictionary plan, the lexicographical file, and the dictionary (Wiegand 1984:14).

3.3.2 What is the structure of meta-lexicography?

Now is the time to explain his structure of meta-lexicography. In his article, it is not easy to find a one-sentence definition of meta-lexicography, but he mentioned that meta-lexicography has four components, i.e.,

> (1) The history of lexicography
>
> (2) The general theory of lexicography
>
> (3) The research on dictionary use
>
> (4) The criticism of dictionaries (Wiegand 1984:15)

One of the characteristics of the Wiegand era is that Wiegand analyzed meta-lexicography as having the above four components. Wiegand specified the general theory of lexicography as having four constituent theories, i.e.,

(1) The general section
(2) The theory of organization
(3) The theory of lexicographical research on language
(4) The theory of the lexicographical description of language (Wiegand 1984:15)

Wiegand explored each constituent deeply, and fully explained the components of each theory and produced the following diagram except Theory B, which is the theory of organization.

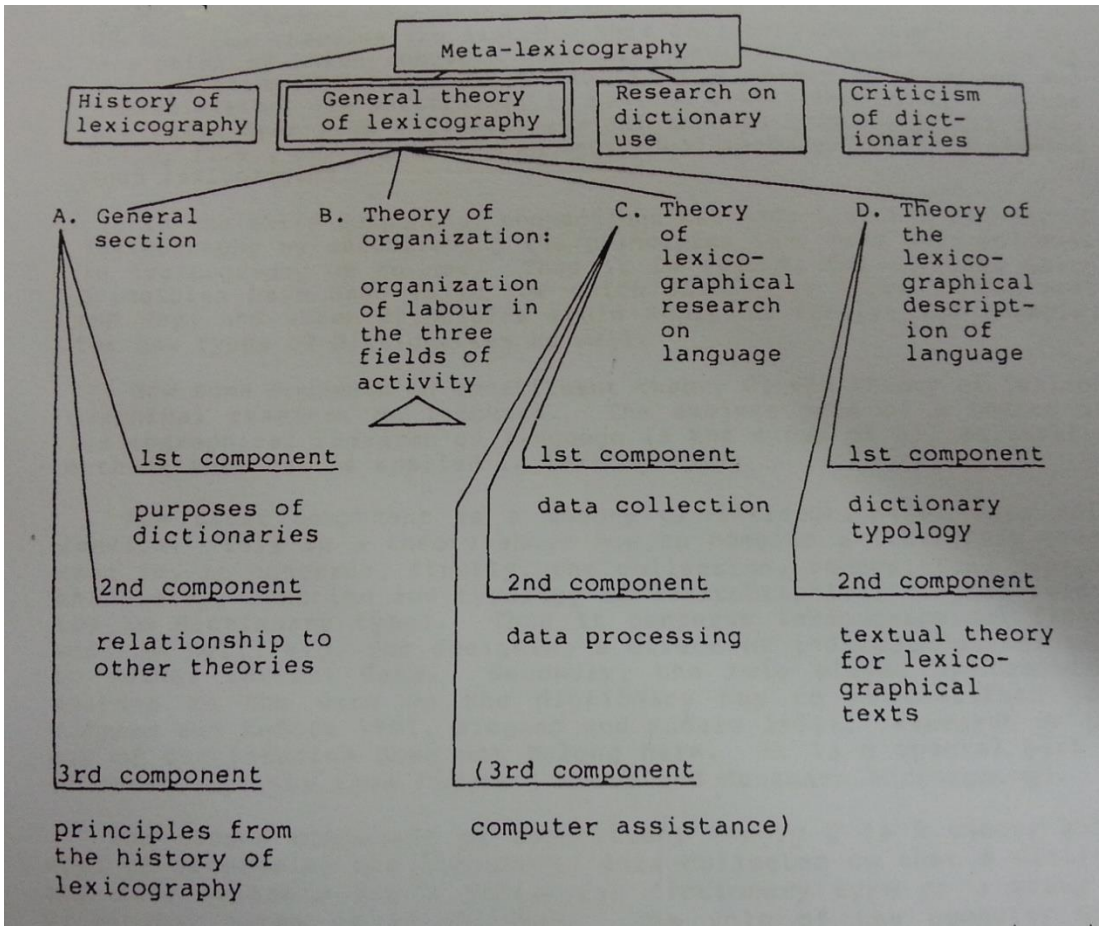


Figure 2. Meta-lexicography (Wiegand 1984:15)

And also, Wiegand (1983:15-16) characterized each individual constituent theory and its components as follows:

A. General section:		Relationships between 1. society 2. other theories 3. the history of lexicography
	First component:	General purposes for monolingual and bilingual dictionaries are derived from the communicative and cognitive needs of the society.
	Second component:	The connections with other theories or constituent theories are listed.
	Third component:	The connections are made with the history of lexicography by establishing the principles that have been followed in lexicography up to now.
C. Theory of lexicographical research on language		The subject area of a theory of lexicographical research on language is the class of all scientific methods that can be applied in lexicography
	First component:	A theory of lexicographical data collection. 1. the collection, composition, representativity, function and typology of lexicographical corpora relative to dictionary types. 2. the role played by secondary sources in the work on the dictionary has to be clarified.
	Second component:	A theory about the ways of processing the linguistic data that was collected.
D. Theory of the lexicographical description of language		The subject of a theory of lexicographical description of language is the classification of all the presentations of the results of linguistic lexicography as texts about language.
	First component:	A dictionary typology and its rationale.
	Second component:	The structure of lexicographical texts.

Table 4. Constituents and components of the General Theory of Lexicography (Wiegand 1984:15-17)

3.3.3 Wiegand on the user-perspective dictionary

In the Wiegand era, there was a strong support of the needs and the reference skills of the target users of dictionaries (Gouws and Prinsloo 2005:5). Wiegand was the first lexicographer to define the dictionary as a utility product, produced to satisfy certain human needs (Bergenholtz and Tarp 2003:178). He started to use the term "genuine purpose" to explain that the dictionary as a utility product has a purpose.

> Wie alle Gebrauchsgegenstände, so haben auch Nachschlagewerke genuine Zwecke. (Wiegand 1998:52)
>
> [Like all utility products, reference works also have a genuine purpose.]

He explained that this genuine purpose is to enable a potential user to retrieve the information, which the user wants to find.

> Ein Sprachwörterbuch ist ein Nachschagewerk, dessen genuiner Zweck darin besteht, daß der ein potentielle Benutzer aus den lexikographischen Textdaten Informationen zu sprachlichen Gegenständen gewinnen kann. (Wiegand 1998:53)

> [A language dictionary is a reference work whose genuine purpose is to enable a potential user to retrieve information about linguistic objects from its lexicographic data.]

In his book (Wiegand 1998), he acknowledged the practical value of the user-perspective, and he includes comprehensive discussions of research methods with regard to dictionary use. Through his typical tests and experiments, various aspects of dictionary use could be investigated.[24]

3.4 The theory of lexicographic functions

3.4.1 History of the theory of lexicographic functions

As Gouws and Prinsloo (2005:7) mentioned, the notion of lexicographic functions is nothing new. There are a few scholars who have developed the function theory. Ponsonby A. Lyons is the person who introduced the concept of user perspective. Tarp (2008:15) said that this was to revolutionize meta-lexicographical thinking in the centuries to come. He quotes Hausmann about Lyons as follows:

> A Dictionary of Language should contain all the words which may be reasonably looked for in it, so arranged as to be readily and surely found and so explained as to make their meaning and if possible their use clear to those who have a competent knowledge of the language or languages in which the explanations are given (Hausmann 1989:89).

Lev V. Scerba is the first person who outlined the principles of a future general lexicographical theory (Tarp 2008:17). On the basis of his experience in working on a Russian-French dictionary, he wrote an article in 1940, but this article did not become known to the world until its translations into German (1982) and English (1995) were published. His reputation started based on his insight of differentiating defining dictionaries and translating dictionaries. Scerba defined both dictionaries:

> Defining dictionaries are intended in the first place for native speakers of a language. Translating dictionaries arise in response to the need to understand texts in a foreign language. (Scerba 1940:338)

Through his personal experience, he found that translating dictionaries did not provide any help to students to learn the language in the process of translation, but just helped them to guess the sense of the word in its context (Tarp 2008:19, Scerba 1940:340). After acknowledging this truth, he suggested to the students that they discard the use of translating dictionaries:

[24] This information was given by R. H. Gouws in a private discussion on April 20, 2016 at Stellenbosch.

> In the light of all this, any true pedagogue advises students to discard translating dictionaries as soon as possible and switch to the defining dictionary of the foreign language. A translating dictionary, then, is only useful for beginning foreign language students (Scerba 1940:341).

Scerba's main idea was that, when students learn a foreign language, they have to learn the language by the system of that language, instead of using the system of their mother tongue. For this reason, Scerba suggested translating the foreign-language, monolingual defining dictionary into the user's mother tongue. In this regard, Tarp (2008:19) conveys Scerba's strong position:

> Scerba concludes that until such dictionaries have been produced, traditional translating dictionaries from a foreign language into the user's mother tongue will remain a malum necessarium, but to overcome this situation as soon as possible he suggests specifically that Larousse's monolingual French dictionary should be "translated" into Russian (Tarp 2008:19).

There were some people, who applied Scerba's principles and tested them with minor changes. For example, Mikkelsen (1992:27) used Scerba's principles and developed them into general principles for a Russian-French dictionary designed for Russian users (Tarp 2008:20). However, in his general principles, Mikkelsen does not agree with Scerba's principle for an explanatory dictionary:

> Provide a translation, not an explanation, that will, in the appropriate grammatical form, fit into a correct French sentence which has been translated from a Russian sentence…. (Mikkelsen 1992:27)

Duda also said that in Scerba's principle for an explanatory dictionary, the meaning-based definition appears to be much more difficult (Tarp 2008:21):

> The user is apparently able to understand that analysis of a word's meaning as it is given in the definition in a monolingual dictionary. It appears to be much more difficult for the user to state the meaning based on a given definition, i.e. to perform lexicalization (Duda 1986:13).

Even though there was some criticism, Scerba's basic idea that a learner's dictionary should be designed based on specific, didactic foreign-language ideas, was developed into a user-oriented principle after ten years by Tarp (2008:21).

Hausmann, a German lexicographer, wrote *Introduction to the Use of New French Dictionaries* in 1977 (Tarp 2008:21-25). In this book, he divided foreign-language dictionaries into two main categories: learning dictionaries and consultation dictionaries. The differences between these two terms are summarized in the table below:

	Learning dictionaries	**Consultation dictionaries**
Designed for	Partial or complete processing	Consultation
With a view of	Global issues	Punctual issues[25]
Sample questions	• How is the word written? • How is it pronounced? • Is the noun masculine or feminine? • How is the word marked?	• How is the French vocabulary structured overall? • In which word families, are synonym and antonym structures found?
Subcategories	1. Primary dictionaries: thorough processing 2. Secondary dictionaries: partial processing	1. Writing dictionaries for writing foreign language texts 2. Reading dictionaries for reading and understanding foreign language texts

Table 5: Differences between Learning dictionaries and Consultation dictionaries (Hausmann 1977:144)

The typology of Hausmann is more developed than Scerba's principles in regard to a user-oriented dictionary. Furthermore, his global and punctual issues gave great insight to Tarp's function theory which followed (Tarp 2008:24).

Finally, Herbert Ernst Wiegand, mentioned in 3.3, has a complicated love/hate relation with function theory (Tarp 2008:28). In the beginning, function theory shared a common foundation with Wiegand, but later it was differentiated from Wiegand and developed in its own way. Wiegand and Tarp have two common foundations: 1) both of them declared that lexicography should be independent from the other sciences such as linguistics, cf. 3.3.1. 2). Both understand dictionaries to be objects of use to satisfy specific human needs (Tarp 2008:29).
One of the important concepts in Wiegand's theory is that of the genuine purpose of dictionaries. Wiegand defines the genuine purpose as follows:

> At the highest level of generalization, there is only one genuine purpose which all dictionaries have. It can be defined thus: The genuine purpose of a dictionary is that it can be used to acquire particular information about language or the non-linguistic world from lexicographical data belonging to certain data types, excluding information about the dictionary used (Wiegand 1987:200).

25 "Punctual consultation" refers to a consultation where the user wants immediate assistance, i.e. a quick solution to a specific problem, e.g. the meaning of a given word or a specific translation equivalent. For such a user, the more general aspects of the lexicographic presentation of the dictionary are not that important at that stage (Gouws in private Email response on 21 Aug., 2016).

From this genuine purpose of dictionaries, Wiegand developed three typologies (Tarp 2008:29), i.e. 1) a typology of dictionary users 2) a typology of situations in which dictionaries are used, and 3) a typology of so-called search questions. And, furthermore, he formed some concepts from these typologies such as trained users, potential users, dictionary addressees, and exemplary users (Tarp 2008:30, Wiegand 1987:218). Among these concepts, the potential users are one of the most important aspects in function theory, because they inspired the formation of the function theory. On the other hand, Tarp stated that Wiegand failed to further develop this important concept:

> Even though he introduces the concept of potential users which function theory subsequently adopted, he does not develop this important concept—which never achieves any major significance in his lexicographical theory (Tarp 2008:30).

3.4.2 The main focus of the theory of lexicographic functions

In 1987, a group of people in the Aarhus School of Business in Denmark, which included Sven-Olaf Poulsen and Henning Bergenholtz, developed a lexicographical theory and published *Leksikografi på HHÅ. Udvikling og perspektiver* [Lexicography at the ASB, Development and Perspectives] (Tarp 2008:33). The common theme of this group in regard to lexicography was 'dictionary users as an object of research'. As in 3.4.1, Bergenholtz, Tarp and their team started their function theory on the basis of Wiegand's theory which included dictionary users, genuine purpose of dictionaries, and potential users (Tarp 2008:34). However, as time went by, the gaps between their understandings and Wiegand's ideas became wider and wider. When Tarp (2008:40-41) explained the basic difference between Wiegand's general theory and the theory of lexicographical functions, he said that the latter shifts the focus from actual dictionary users and the dictionary usage situation to potential users and the social situations in which they participate. He added the reasons for this shift, one of which is stated as:

> Dictionaries are basically an answer to specific types of need registered in society among specific types of user in specific situations…. As a result, the useful value of dictionaries must be seen in relation to these needs, instead of being determined phenomenologically based on whether the dictionary user's consultation is successful or not…. (Tarp 2008:40).

So, this substantial research into the user's need is always connected to a specific person located in a specific situation in which the needs concerned have arisen (Tarp 2008:41).

3.4.3 Main elements of the theory of lexicographic functions

In function theory, according to Tarp (2008:43) there are four elements, three extra-lexicographical elements, and one intra-lexicographical. The three extra-lexicographical elements are: potential users, user situation, and user need. And the intra-lexicographical element is: assistance. However, it is not easy to discover the user's needs and situation. Since the needs are related to the user's situation, Tarp used a deductive procedure for discovering the situation in which the needs arise.

3.4.3.1 Communicative and cognitive situations

There are two fundamentally different situations: communicative situations and cognitive situations (Tarp 2008:44-45). First, the cognitive situations arise from the need to gain new knowledge such as,

> 1) while reading—the sudden wish to learn more about a given question;
> 2) while writing—the need to know more about a given topic in order to finish a text;
> 3) during discussions, or when entering wagers about specific issues;
> 4) during processes in the subconscious—the sudden desire to examine something;
> 5) during dictionary consultation—the desire to know more about a specific topic;
> 6) in relation to specialized translation and interpretation tasks;
> 7) in relation to a teaching program;
> 8) in relation to a course of study (Tarp 2008:45)

Tarp (2008:46) explained that the first five situations mentioned above are associated with punctual issues of users, and the rest of the items[26] are also punctual, but are leading to a specific dictionary consultation, connected to a global issue[27]. From this analysis, the following four types of cognitive needs can be produced: 1) encyclopedic knowledge of linguistics (LGP[28]), 2) specialized linguistics (LSP[29]), 3) general, 4) cultural and special-specific need.

Secondly, the communicative situations are more complicated. Let's first look at the possible processes of communications from one person to another (Tarp 2008:47-53), for the first model,

[26] Specifically, the seventh and eighth situations, and, to some extent, the sixth.
[27] These punctual and global issues come from Hausmann's theory (Cf. 3.4.1).
[28] Language for General Purposes
[29] Language for Special Purposes

one person (sender) produces a text which is then received by another person. This model is under the assumption that the languages of the sender of the text and the recipient of it are the same as their mother tongue. In this model there are three levels, i.e. player, process, and text level. A simple communication model revealing lexicographically relevant situations in which the potential user is involved is shown below (Tarp 2008:47):

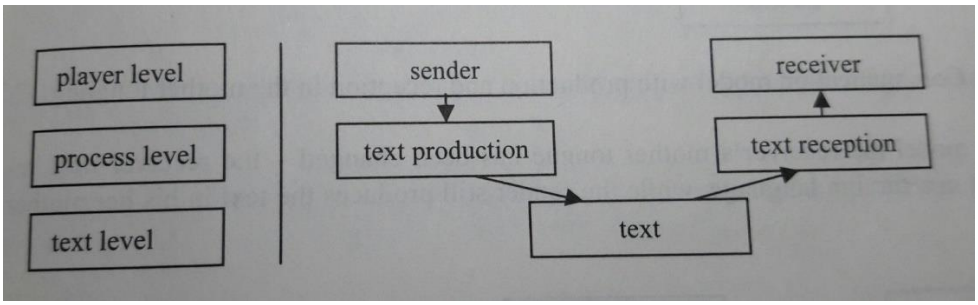


Figure 3. Simple communication model revealing lexicographically relevant situations

In this model we can find two lexicographically relevant situations (1) and (2):

1) production of the text in the mother tongue
2) reception of the text in the mother tongue

And then if we suppose the languages of the sender and recipient are different, and a translator is added between the sender and recipient, lexicographically relevant situations will be more complicated by the person who will translate the text. The sender or the recipient? (Tarp 2008:48).

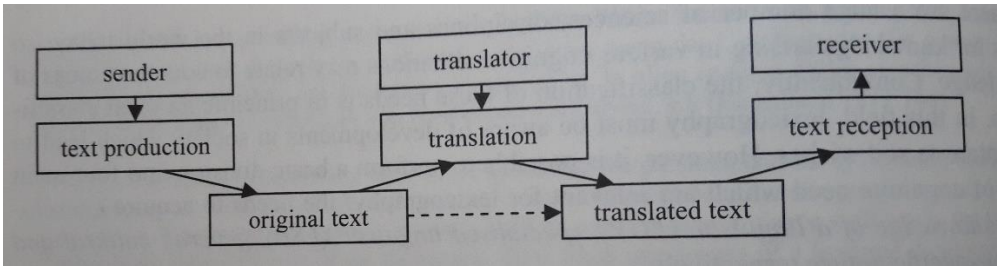


Figure 4. Simple translation model revealing lexicographically relevant situations

In this model we can find five lexicographically possible relevant situations (3) - (7):

3) production of the text in the foreign language
4) reception of the text in the foreign language
5) translation of the text from the mother tongue into the foreign language
6) translation of the text from the foreign language into the mother tongue
7) translation of the text from one foreign language into another

And then again if we suppose the case that a proofreader for checking the text or a teacher for marking it will be added, and also the languages of the sender and proofreader or teacher are different, lexicographically relevant situations will be more complicated as given below (Tarp 2008:52-53):

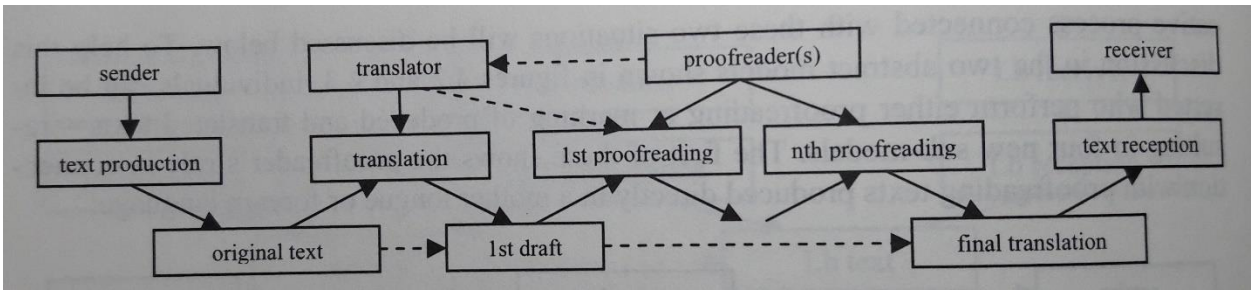


Figure 5. Model for communication in connection with proofreading translated texts

In this model we can find another five lexicographically possible relevant situations (8) - (12):

8) proofreading or marking of the text produced in the mother tongue

9) proofreading or marking of the text produced in the foreign language

10) proofreading or marking of the text translated from the mother tongue into the foreign language

11) proofreading or marking of the text translated from the foreign language into the mother tongue

12) proofreading or marking of the text translated from one foreign language into another

Since it is not easy to find the user's needs and situation, Tarp analyzed 8 cognitive situations, and 12 communicative relevant situations by a deductive procedure for discovering the situation in which the needs arise.

3.4.3.2 Needs of potential users

As stated in 3.4.2, even though function theory uses the same term, potential users, as Wiegand does, the focus has shifted from the actual dictionary to the potential users and the social situations. To discover the exact needs of the users, we have to study all types of users and various types of user situations. For this purpose, Tarp (2008:56-57) sub-divided the types of needs into two categories: primary and secondary user needs. Primary needs are the needs for information leading to a dictionary usage situation:

- information about the mother tongue
- information about the foreign language
- information about specialized language in the mother tongue

- information about specialized language in the foreign language
- comparative information about the mother tongue and the foreign language
- comparative information about specialized language in the mother tongue and the foreign language
- general cultural information
- information about the culture in a specific language area
- information about a specific subject or science
- comparative information about a subject in the national culture and the foreign culture (Tarp 2008:57)

Secondary needs include needs for information and needs for instruction and education:

- general education in lexicography
- general instruction in dictionary usage
- information about the specific dictionary
- instruction in the use of the specific dictionary (Tarp 2008:57)

3.4.3.3 Potential user's lexicographical qualifications

For the specialized dictionary, the qualification of the potential user can determine the extent of lexicographical data. Traditionally the potential users of specialized dictionaries were divided into laymen and experts (Tarp 2008:78). Rasmussen divided them into three: laymen, semi-experts and experts. She understood that there are two elements such as information needs and specification needs. For laymen, information needs are high, but specification needs are low; whereas for the experts specification needs are high, but information needs are low (Tarp 2008:78-79).

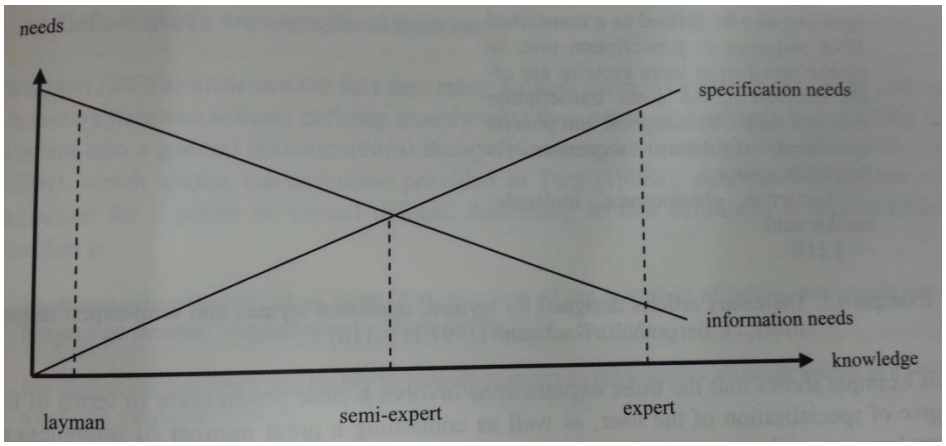


Figure 6. Relationship between specialized knowledge and types of need (Rasmussen 1998:142)

Tarp introduced the following three questions to define the user's qualification:

1) How much do users know about lexicography?
2) What general experience of dictionary usage do they have?
3) What specific experience do they have with a specific type of dictionary? (Tarp 2008:55-56)

3.4.4 Influences of function theory on dictionaries for endangered languages

In function theory, what are of most interest are the various types of user and user situations. If we think about the environment of the endangered languages, there could be one more factor, especially the situation of the language itself. In the general lexicography studies, this latter factor was not mentioned at all. I think the reason is that historically the lexicography studies were designed for major languages, which are fully developed, not for minority languages, which are under development or just starting to be analyzed. Because the language under discussion is itself not so fully developed, the possible types of dictionaries to use in this situation would be:

- dictionary or glossary for teaching the alphabet or orthography
- dictionary for language standardization
- dictionary for different dialects
- dictionary for language documentation

Since most of the dictionary users, in the case of endangered languages, are probably uneducated people, the qualification of the user is another topic to be considered in lexicography. As Rasmussen (1998) explained in Figure 3, information needs are more necessary than specification needs for uneducated mother-tongue speakers. However, if the dictionary is bilingual or multilingual, the users and their situations are much more complex. So the functions of the envisaged dictionary will be as follows;

First, the users, whose qualifications in lexicography are likely to be relatively high, higher than the semi-expert level of Rasmussen's chart, are classified below:

- educated mother-tongue speakers
- the same nationality, but speakers of other languages
- foreign speakers

Second, the user situations will be as follows:

- special kind of reception in the mother tongue
- translation-related reception in the foreign language
- special kind of production in the foreign language

3.5 Chapter summary

For field lexicographers, there are enough data, but their work does not stand on firm theoretical foundations. It is not necessary for all of them to be theory specialists. It will take a long period of time for them to become familiar with lexicographic theory. In this chapter, I discussed the history of meta-lexicography, and explained each lexicographic theory for these lexicographers. Lexicographers should always keep in mind the genuine purpose of dictionaries and also the target users' reference skills. They have to determine which type of dictionaries is best and which functions are needed for each endangered language. This demands the formulation of a dictionary conceptualization plan in the early stages of the lexicographic process. Without this comprehensive planning, the process will be long and cumbersome. I realize that there are more theories than the ones I have explained. In this chapter, I tried to focus only on a few theories that are relevant for the compilation of a dictionary for endangered languages.

Chapter 4. Data collection for endangered languages

4.0 Introduction

The purpose of this chapter is to describe methods of data collection, especially for endangered language groups, in order to compile their dictionaries. Data collection for endangered languages is quite different from that for languages, which have already been developed linguistically. Most endangered languages have not yet been surveyed to discover whether the language form under consideration is the standard dialect or not. Before making an extensive collection of lexical items, the varieties of the language and the dialect map should be studied. Otherwise, it will take years to identify and sort out the mixture of varieties in the collected data.

In this chapter, the prerequisite language study will be mentioned first, then two methods of data collection will be explained, i.e. building text corpora and collecting data using semantic domains. These two methods complement each other. The gaps in the vocabulary of the language, which are not filled using the text corpora, will be filled by the words elicited according to semantic domains.

4.1 Prerequisite language study

For a lexicographer, who is compiling a general language dictionary, an accurate understanding of the language itself is very crucial. In particular, he or she has to know about the varieties of the language. Personally, when I compiled the Nepali-Korean dictionary, the varieties of Nepali were a major problem (Lee 2003:81). A sociolinguistic survey of Sherpa will be the most likely way to determine the most generally accepted variety of this language. This survey exposes not only the relationship between Sherpa and the neighboring languages within the same language family, but also the relationship between its different dialects. The main role of this survey will be to find out the standard dialect from among the varieties[30]. In this subsection, the sociolinguistic survey of the Sherpa language will be presented as an example (Lee 2003) before considering the two methods of data collection.

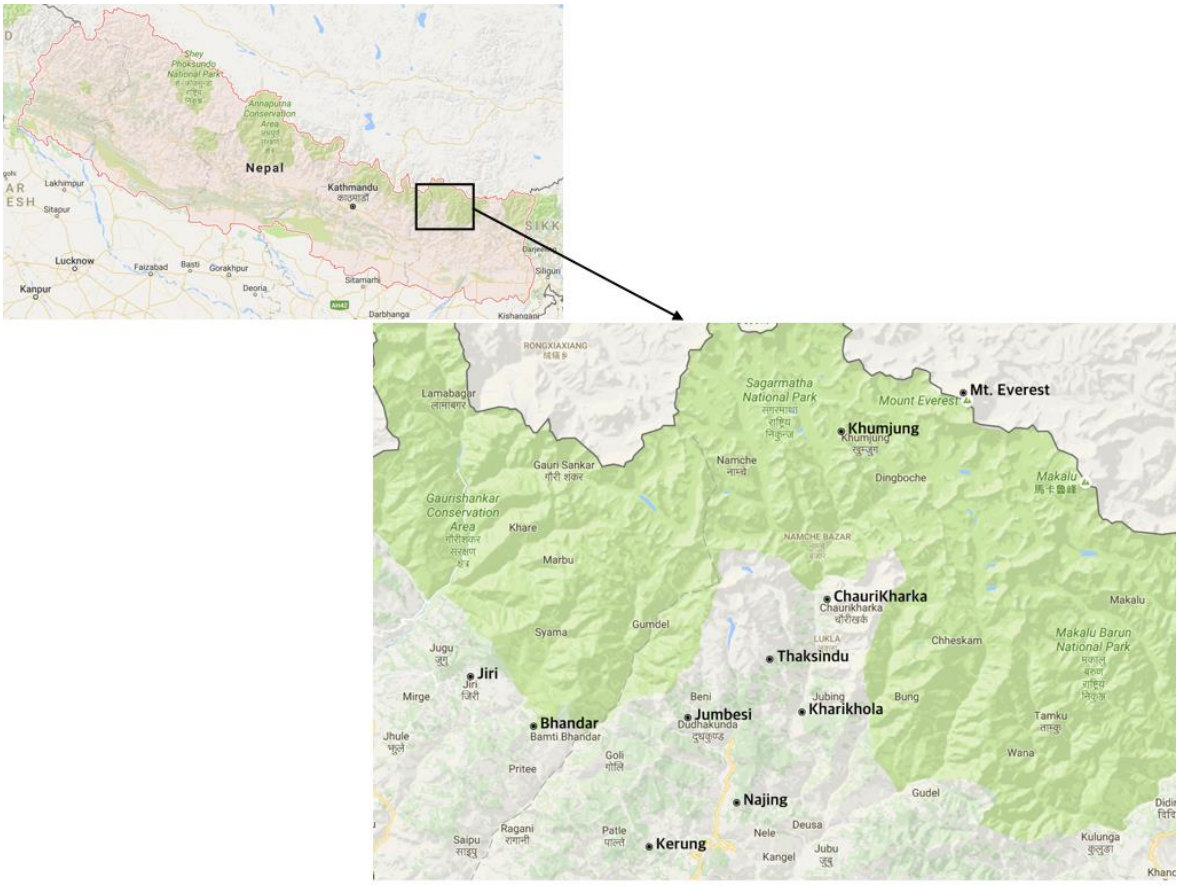
[30] This sociolinguistic survey was done with the help of Troy Bailey.

4.1.1 Lexical similarity comparison to discover the language varieties

The purpose of the lexical similarity comparison test is to find out the relationship among the neighboring languages and dialects. This method compares a list of 240 words (see the 240 word-list in Appendix 1) that has been elicited from each of the target languages and dialects. This method was borrowed from Blair (1990), and was adapted in some parts as needed. Through this comparison, the similarity among the languages and dialects could be calculated (Lee 2003:84). The following subsections are sample test results from the Sherpa language (Lee 2003:85-87). The names of the two neighboring languages and the nine varieties of Sherpa that were compared are as follows:

- Nepali: national language of Nepal, an Indo-Aryan language.
- Lhasa Tibetan language, in the same language tree as Sherpa under Central Tibetan (Hale 1970)
- Western dialects of Sherpa: Jiri, Bhandar
- Southern dialects of Sherpa: Jumbesi, Kyerung, Najing, Thaksindu
- Northern dialects of Sherpa: Kharikhola, Chaunrikharka, Khumjung

The following maps are the map of Nepal, followed by the map of the Sherpa language speaking area. Nine places, where the test was done, are marked on the second map.



Map 1: The map of Nepal and nine Sherpa speaking villages, which were tested[31]

4.1.1.1 Result (1): Lexical similarity percentages

The lexical similarity percentages calculated in this survey are shown in Figure 7.

Nepali										
3	Jiri									
3	88	Bhandar								
3	84	87	Jumbesi							
3	86	87	86	Kyerung						
3	87	89	90	92	Najing					
3	85	89	88	85	89	Thaksindu				
3	81	85	82	83	86	89	Kharikhola			
3	80	83	82	81	85	85	91	Chaunrikharka		
4	77	82	78	79	82	84	96	87	Khumjung	
1	32	33	30	33	32	32	33	34	35	Lhasa Tibetan

Figure 7. Matrix arranged by geographical position (Lee 2003:86-87)

[31] Source: http://maps.google.com, edited by Daewon Kim

This matrix shows that Nepali, of which the similarity is less than 3%, is from a completely different language family as Sherpa. Lhasa Tibetan is in the same language family as Sherpa, but the percentages of similarity are less than 35%, which means that it is a different language from Sherpa. Finally, among the Sherpa varieties, one fact discovered is that the percentages between the Western (Jiri) and Northern (Khumjung) dialects are 77%[32], which is the lowest among the Sherpa varieties (Lee 2003:86).

4.1.1.2 Result (2): Lexical differences

The comparison of the nine varieties of Sherpa shows that there are lexical differences among these three dialects (Lee 2003:86). See the sample in Table 6.

Gloss	West	South	North
SAND	pepsok^{1}	pema ^{1}	p^{h}ema^{1}
SPIDER	bəldzjaŋ^{1}	bəldzjaŋ^{1}	rhilkoŋgoŋ^{2}

Table 6. Lexical differences (Lee 2003:85)

4.1.1.3 Result (3): Phonological differences

Voiceless aspirated stops in Khumbu (North) correspond with simple voiceless stops in South and West (Lee 2003:86). See the examples in Table 7.

Gloss	West & South	North
DAUGHTER	pum^{1}	p^{h}um^{1}
WHEAT	ʈa^{1}	ʈ^{h}a^{1}
MOUNTAIN (covered by snow)	kaŋri^{11}	k^{h}aŋri^{11}
EGG	tsemendok^{211}	ts^{h}emendok^{211}

Table 7. Phonological differences (Lee 2003:85)

Another difference is the absence of the voiceless velar /k/ of the North in the word final position of South and West dialects, and furthermore '–wu' is added (Lee 2003:86). Examples are shown in Table 8.

Gloss	West & South	North
KNIFE	ʈ^{h}iwu^{12}	ʈ^{h}ik^{2}

[32] In Figure 7, 77 % was obtained by reading the number where the vertical column and the horizontal row for those languages intersect.

MONKEY	rhiwu^2	rhik^2

Table 8. The absence of voiceless velar (Lee 2003:86)

Along with other evidence found in the sociolinguistic survey report, these phonological differences confirm that there are three dialects of the Sherpa language (Lee 2003:87).

4.1.2 Dialect intelligibility study to discover the standard dialect

After confirming that there are three dialects of the Sherpa language, for the purpose of compiling a Sherpa dictionary, a process was needed to find the standard dialect of Sherpa. Which dialect should be the basis of the Sherpa dictionary? The lexical items of the other two dialects will not have their own articles, but be treated as guiding elements toward the standard dialect.

A big question regarding the standard dialect was: Who would choose it? The best way would be to ask the language society itself to decide. However, this is never an easy way. Personally, while I was working on the Sherpa dictionary as an expat specialist, the decision concerning the standard dialect was the most difficult process. Even though I had waited for 30 years, still the Sherpa people had not decided on a dialect to be regarded as the standard Sherpa dialect. Furthermore, the government did not have any desire to be involved in dialect struggles. At this point, the dialect intelligibility test was a second option in this situation.

This test is based on a linguistic approach, not on a political or regional or historical approach. Casad (1974) developed the Recorded Text Tests (RTTs) to see how well the speakers of one dialect could understand the other dialects. The results of the dialect intelligibility test using RTTs within the Sherpa language are shown in Table 9.

	SUBJECTS		
	West	**South**	**North**
Western text	AV=95% N =10	Test not done	Test not done
Southern text	AV=96% N =10	AV=97% N =11	AV=95% N =10
Northern text	AV=81% N =10	Test not done	AV=96% N =10

Table 9. Result of Intelligibility test (AV= average, N= number of subjects) (Lee 2003:88)

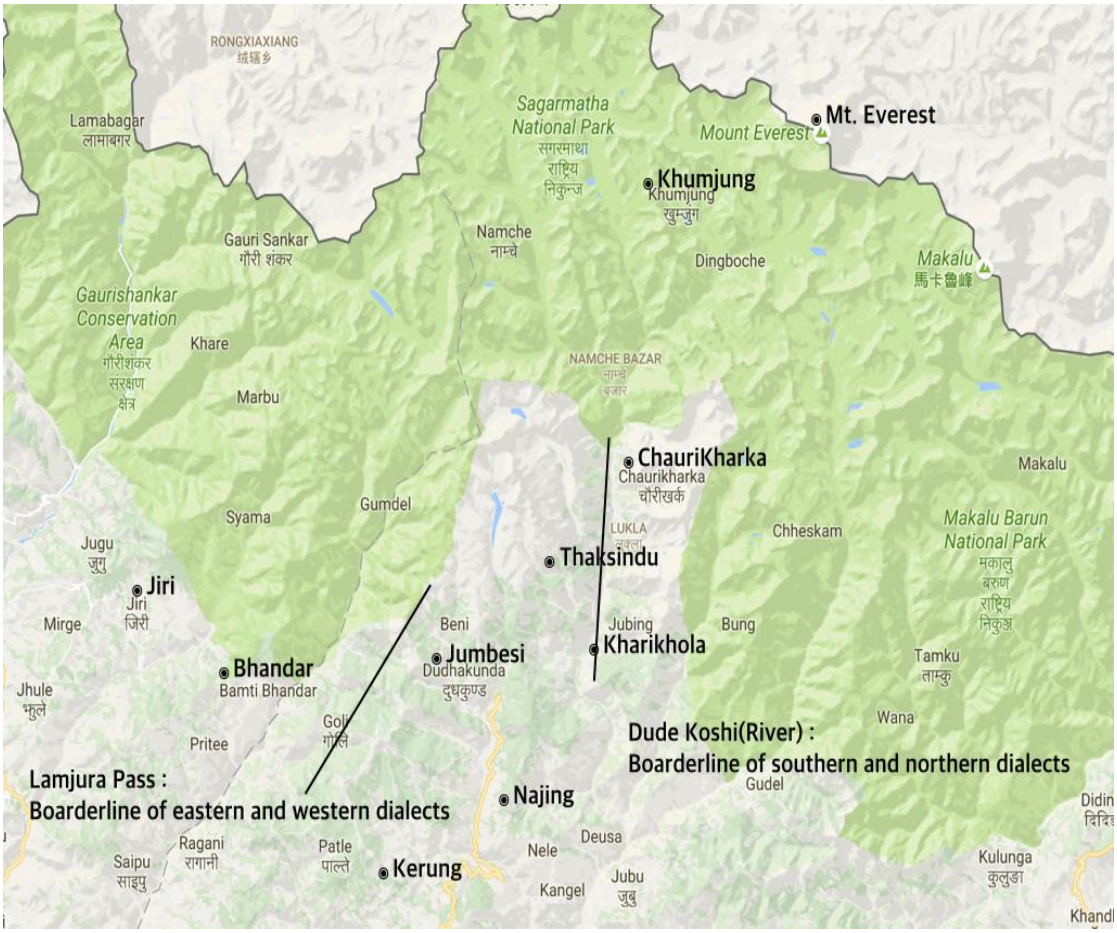
Even though the test was not done throughout the whole area, the results of this test give enough data to form the following conclusions (Lee 2003:88):

- The understanding of the Western people of the Northern text was lower, so that there is a comprehension problem between the Western and Northern dialects. This also aligns with the fact that the lexical similarity percentage (77 %) between Jiri (Western dialect) and Khumjung (Northern) was very low in 4.1.1.1.
- The understanding of the Southern people of both the Western and Northern texts was reasonably high, so the Southern dialect could be a bridge or understandable dialect to both the Western and Northern dialects. This means that the Southern dialect could be considered the standard dialect of the Sherpa language.

On the basis of this conclusion, the Sherpa language map could be drawn. The border between the Western and Southern dialects is the Lamjura Pass[33], and the border between the Southern and Northern dialects is the Dudh Khosi River[34]. This language map is a very important tool for obtaining the text collection needed for building the corpora mentioned in the next section. For each text collected, the region of the resource person should be recorded, so that it will be clear to which dialect the text belongs. The texts should be collected from all three dialects separately, but for the forms to be included in the dictionary as lemma signs it is better to collect more texts from the Southern dialect. Sherpa language map is as follows:

[33] 3,400 meters in altitude

[34] Dudh Khosi River, beginning from the Southern part of Mt. Everest, flows toward the South and divides the Northern and Southern dialect speaking areas at Kharikhola. From there, the eastern part is the Northern dialect speaking area, and the western part is the Southern dialect speaking area. There is no access between the Western and Northern dialect speaking areas, as it is full of high mountains along the Dudh Khosi River.



Map 2: The Sherpa language map [35]

4.2 Data collection by corpus building

After the completion of the dialect map, the dictionary compilation process is preceded by the collection of written and spoken texts (Gouws and Prinsloo 2005:21). Newell explains what a text is as follows:

> A text may be composed in written form or may be a transcription of speech, usually recorded by means of a cassette recorder. The sample might range in length from a single sentence to the size of an entire volume. (Newell 1995:27)

By whatever instrument the language is recorded, the spoken text should be transcribed on computer to be used as a database for the compilation of a dictionary. Gouws and Prinsloo call it an *electronic corpus*, which they define as follows:

[35] Source: http://maps.google.com, edited by Daewon Kim

> An electronic corpus can be defined in an oversimplified way as a computerized collection of texts. (Gouws and Prinsloo 2005:21)

The reason for building a corpus before actually compiling a dictionary, as Kennedy explains, is to be the basis for the analysis and description of the structure and use of the language (Kennedy 1998:60). There is a significant difference between a dictionary based on a corpus of the "living language" and one that is not (Newell 1995:28).

4.2.1 Limitations in building a corpus for the endangered languages

Building a corpus for endangered languages is quite different from a large project such as the compilation of a dictionary of the English language, which requires a corpus of several million words (Newell 1995:27). McEnery and Ostler (2000) divided the languages of the world into four broad types:

1. Official majority languages (e.g., English in the UK, Portuguese in Portugal).
2. Official minority languages (e.g., Welsh in the UK).
3. Unofficial languages (relatively large, e.g., Kurdish in Turkey and relatively small, e.g. Sylheti in the UK).
4. Endangered languages (e.g., Guugu Yimidhirr in Australia).

They added that for types 1 and 2 it will be easy to build the corpora with the government's support and with strong financial support. However, type 3 will suffer from lack of official recognition and state funding. In the case of endangered languages, obviously–very few speakers will produce very little material. In some cases, the language will also be suppressed (McEnery and Hardie 2012:12).

4.2.2 The design of the corpus for endangered languages

In this subsection, the type of corpus, an issue of the quality and the quantity of the corpus will be mentioned. 'Representative' and 'balanced' are the quality issues, and size of the corpus is the quantity of text.

4.2.2.1 The type of the corpus

In the corpus for endangered languages, the type will be, first of all, be *static* rather than *dynamic*, because the materials collected in the endangered languages cannot give enough evidence to monitor the changing patterns of usage over time (Kennedy 1998:61). And secondly, the type

will be *general*, so that the corpora will be used for the linguist to analyze and seek answers to particular questions about the vocabulary, grammar, or discourse structure of the language. This general corpus is typically designed to be balanced, by containing texts from different genres and domains (Kennedy 1998:19-20). And thirdly, the type will be *synchronic* rather than *diachronic* (Kennedy 1998:22), because in most of the endangered languages there are few written or oral corpora, which were produced in the past, so that we do not have enough data to compare the changes over a period of time.

4.2.2.2 The quality of the corpus: the representativeness and balance issue

The quality of a corpus directly influences the quality of a dictionary, because many aspects of the lexicographic treatment are based on the analysis of this corpus. This is the reason why the corpus should be representative by including data from all spheres of the speech community, different registers and dialectal variations. And also it should be balanced by containing texts from different genres and domains of use including spoken and written, private and public (Gouws and Prinsloo 2005:23, Kennedy 1998:20). Issues associated with how to make a corpus representative or balanced or able to be used for comparative purposes are essentially issues concerning the quality of a corpus (Kennedy 1998:66). The representativeness issue is not easy to define. There are still different ideas about 'Representative of what?' (Kennedy 1998:62). McEnery and Hardie (2012:10) commented that the measures of balance and representativeness are matters of degree. Even though the representativeness is not accomplished completely, we have to try to get as high a degree of representativeness as possible in the process of building the corpus.

There will be three ways to increase the degree of representativeness. First, a corpus should be restricted to a single dialect (Newell 1995:35). For the general corpora, it will be good to collect texts from the whole dialect-speaking area to see the variants, but each corpus should be recorded by the name of the specific dialect and location. If all dialects are included in the corpora from the beginning without documenting the dialect and location, phonological and lexical differences among the dialects will be mixed together and it makes all corpora confused (see 4.1.2).

Secondly, the choice of the native speakers, who will give the spoken texts, is crucial. Newell gives the aspects to be considered with regard to the native speakers:

1) A fluent speaker of the language
2) An outgoing person
3) A well-respected adult
4) One immersed in the language and culture and comfortable in the language and culture, not a recent immigrant
5) A person with a keen interest in both the object language and the culture
6) One with a good network of social relationships within the community (Newell 1995:29-30)

Thirdly, balanced text collections from various genres could make the degree of balance higher. First of all, there are many genres in literature, such as all kinds of stories, fairy tales, legends, poems, hortatory, and instructional materials, experience stories, etc. Newell emphasized that all major genres of the object language should be sampled in order to provide the lexicographer with a majority of the lexemes used in expressing a wide range of cultural experience (Newell 1995:38). For this wide range of cultural lexemes, he used George Murdock's Outline of Cultural Materials Subjects List (Murdock 1987). This list shows more than 700 subjects on all aspects of cultural and social life, so that the lexicographer can elicit qualified cultural corpora. A couple of categories with related topics are as follows:[36]

230 Animal Husbandry
231 Domesticated Animals
232 Applied Animal Science
233 Pastoral Activities
234 Dairying
235 Poultry Raising
236 Wool Production
237 Animal By-Products

360 Settlements
361 Settlement Patterns
362 Housing
363 Streets and Traffic
364 Refuse Disposal and Sanitary Facilities
365 Public Utilities

[36] The whole list is in: www.ingramanthropology.com/uploads/6/8/1/1/6811328/ocm.pdf

366 Commercial Facilities

367 Parks

368 Miscellaneous Facilities

369 Urban and Rural Life (Murdock 1987:2-3)

For the qualified corpus, Newell demonstrated the word-list elicitation approach within a cultural setting (Newell 1995:32). I will cover this topic in 4.3.2

4.2.2.3 The quantity of the corpus: The size issue

This issue is: What is the total number of words necessary for a balanced and representative corpus? It is generally assumed that the 'bigger' the corpus the 'better' it is, so huge corpora were built such as the Collins Birmingham University International Language Database (COBUILD) and the British National Corpus (BNC), of which the corpora are more than several hundreds of millions of words (Gouws and Prinsloo 2005:22). However, if we apply this criterion of number of words to endangered languages, the situation is quite different. Newell reported that a full-time staff person, who is a text gatherer and also a keyboarder, could build a corpus of one million words of text in one year. Moreover, in the first million words of text of Newell's Romblomanon project,[37] 2,000 words occurred only once in frequency (Newell 1995:43). Kennedy experienced the same problem, saying that in a corpus of one million words 40 to 50% of the word types occur only once (Kennedy 1998:67).

For the lexicographer, words from a bigger corpus with high frequency would be preferred, but, if we make a rule to include a word as a lemma which appears at least three times in the corpus, 40 to 50% of the word types would be useless, even though we have collected a corpus of one million words. Newell shows the mutual relationship between the number of unique morphemes[38] and the total number of morphemes in the corpus in Figure 5. Figure 5 has been changed into a graph as in Figure 6 to show the situation visually.

[37] Romblomanon is one of the Bisayan languages of the Central Philippines.

[38] 'unique morpheme' could be understood as 'lexical item'.

Total Morpheme Units in Running Text	1 or more Unique Morphemes	3 or more Unique Morphemes	5 or more Unique Morphemes	10 or more Unique Morphemes
200,000	5,000	2,500	2,000	1,200
400,000	6,800	3,700	2,700	1,700
600,000	8,100	4,800	3,400	2,400
800,000	9,300	5,400	3,950	2,900
1,000,000	10,400	5,800	4,400	3,300
1,200,000	11,000	6,100	4,800	3,500
1,400,000	11,700	6,600	4,900	3,650
1,600,000	12,200	6,950	5,000	3,850
1,800,000	12,800	7,100	5,300	4,000
2,000,000	13,200	7,300	5,400	4,150
2,200,000	13,500	7,700	5,500	4,200
2,400,000	13,700	7,850	5,600	4,250
2,600,000	13,750	7,950	5,650	4,300
2,800,000	13,820	7,975	5,675	4,350
3,000,000	13,900	8,000	5,700	4,350

Figure 8. Chart of unique morphemes occurring in various corpus sizes (Newell 1995:21)

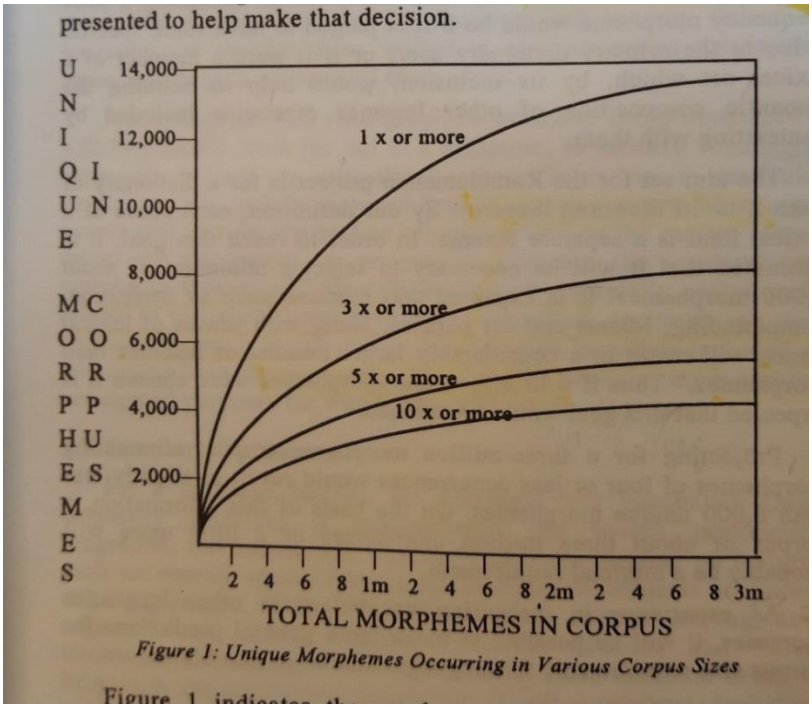


Figure 9. The mutual relationship between the number of unique morphemes and the total numbers of morphemes in the corpus (Newell 1995:21)

Newell explains the relationship between total morphemes in a corpus and unique morphemes in a corpus as follows.

> By eliminating morphemes which occur only once or twice in a three-million morpheme corpus (about 8,000 morphemes), 57.5% of the morphemes remain. If those occurring four times or less are eliminated (5,700 morphemes), 41% of the morphemes remain; and if those occurring nine times or less are eliminated (4,350 morphemes), only 31% remain (Newell 1995:22).

Furthermore, from Newell's data, we find that high frequency morphemes are not easy to find, because the numbers are few. For example, the morphemes occurring ten times or more in a one million morpheme corpus are 3,300, but even though the corpus is expanded to a three million corpus, the morphemes occurring ten times or more are only 4,350. From two extra million corpora, just a little more than one thousand lemmas could be added. [39]

The decision on which lemma will be included or excluded, will depend on the language situation, and on the duration, finances, and man-power of the project. However, if we try to ask what is the proper size of the corpus, Newell's answer is three million word corpora to get 8,000 lemmas (Newell 1995:43), and Prinsloo's suggestion is one million word corpora to get 5,000 lemmas[40] (Prinsloo 2015:299). If we consider it realistically, the size of the corpus will be around one million words, because the materials in endangered languages are quite limited.

4.2.3 Written and recorded spoken texts

Kennedy introduced three ways of capturing text. The first is keyboarding from non-computerized resources, such as handwritten, typewritten, and spoken texts. The second is from computerized texts, such as books, major newspapers, and magazines. If the purpose is commercial, the cost, of course, should be paid for, for using these data. The third is from scanning by the Optical Character Recognition (OCR) method, which is a revolutionary way but often contains a variety of predictable and unpredictable errors (Kennedy 1998:78-80, Gouws and Prinsloo 2005:22). Kennedy suggests that the lexicographer should be sensitive to obtaining the permission of the authors to use their texts both in written and spoken form as follows:

---

[39] Prinsloo shared a similar experience. The expansion of the corpus size from one million to ten million could increase by 5.7% the three-starred words retained (Prinsloo 2015: 289).

[40] Prinsloo mentioned 5,000 lemmas here, because most African publishers restrict dictionaries to a very limited number of pages (Prinsloo 2015: 287).

> Before texts are copied into a corpus database, compilers must seek and gain the permission of the authors and publishers who hold the copyright for a work, or the informed consent of individuals whose rights to privacy must be recognized (Kennedy 1998:76-77).

Newell also advises that caution is needed, because there is an increasing feeling of ownership of language data by the native speakers (Newell 1995:41-42).

As I mentioned in 4.2.1, for the endangered languages, there are many limitations to the opportunities for collecting text. For a dictionary of a major language, the reasonable ratio of written and spoken will be 20% of recorded spoken and 80% of written text (Newell 1995:37), but for endangered languages this ratio is meaningless. The exact situation of endangered languages is that most of the texts would be from recorded spoken text only.

4.2.4 Keyboarding of the text

4.2.4.1 Corpus compiling programs

Before the real process of compiling a corpus starts, one of the most important jobs will be the decision regarding in which computer program to store all the corpora. Each program has its own strong points and weak points. The compiler should consider the best program for the language and for the dictionary type.

Personally I am using Toolbox[41], since the Nepali-Korean dictionary (Lee 1999) and the present Sherpa dictionary, in the process of being compiled had both been started in that program.

4.2.4.2 Text gatherers

Newell advised that, if a non-native speaker has the main responsibility for text collection, there will be disadvantages for the project. One example is as follows:

> One disadvantage is that the text donor almost invariably will have a tendency to give the text gatherer what he or she judges the text gatherer wants to hear. In this respect, the donor often attempts to speak within the context of what is perceived as the donor's cultural background (Newell 1995:29).

When a native speaker is selected as a text gatherer, the following points need to be considered.

1) The dialect of the person should be the same as the dialect of the source language.
2) The person should have a relatively strong degree of identity with the language.

[41] Toolbox was produced by Alan & Karen Buseman. **URL:** http://www-01.sil.org/computing/toolbox/

3) The person needs to have a general understanding of the whole dialect system.
4) The person should be familiar with the spelling system of the language.

4.2.4.3 Keyboarding texts and building a corpus and glossary file

When the text is recorded from either spoken or written materials, each text should have its own file name (Newell 1995:42). The file name consists of the dialect name, Murdock's classification number, and individual numbers under the same classification. For example, in the Sherpa language, the dialect names are composed of one digit: 1 (Northern dialect), 2 (Southern dialect), 3 (Western dialect). Murdock's classification number is three digits (see under 4.2.2.2), for example, 231 (Domesticated Animals), 232 (Applied Animal Science). Finally the individual file number is composed of two digits, when there is more than one file under the same Murdock classification number. One sample file from Sherpa texts is as follows:

File number: 231801[42]
नेपालला दोङ्बु गोवु ताम्डे
नेपालला मी माङ्शोककी दोङ्बु चेतुप ती राङ मिसीन बिरुवा जुवुला सी साङ नासाम मोतोङनोक। तीकि लागी डे डारे युलला बिरुवा चुयिन। चुनी लो च्यिथम्बा चोयी गल। तासाम बेला शोलुक ताङ शीङ चेचे फेन थोकिवी। डे नासामला लो रेरेला मी च्यीककी बिरुवा खाल च्यीक ताङ च्यिथम्बा ति तेरीकी चुसिन ज्युक्थमाला नेपाल लेमु डिवी नोकिनोक। च शीङगी लागी काले मेङ्गिवी। तमा च्यु लुङ साङ लेमु गल्नी मीतिवा साङ छ्ये रिङ्बु डिवी नोकिनोक। दोङ्बु माङ्मु वोसिन छ्यु लेमु डिवी। मी थुङ ताङ राङ लेमु गिवी। दोङ्बु वोसिन डिमा छ्येन्दे शर्सिन ऱ्हीपला देतुप ङ्येकिवी। छर्वु कितुप ङ्येकिवी। दाक्पुला दोङ्बु काङ क्यान गोकिवि सिरुप ताम्डे ति मी तेरीला शे गोकिवी। दाक्पु हाकोवुतुवी हाकमोकोवु मीतिवाला लोप कोकिवी। बिरुवा तिक्पे च्येतुप मेङ्गिवी। च्येसिनाङ दोङ्बु बोम्बु च्ये गोकिवी। दोङ्बु च्यीक च्येसिन बिरुवा सुम जु गोकिवी। दिकी तेन्दोकला दाक्पु तेरीकी नासाम तोङ गोकिवी। दाक्पी खाङ्बी गारिगुरीला बिरुवा जु गोकिवी।

[Translation:
A story for the need of trees in Nepal
In Nepal many people cut trees only, but they don't think of planting saplings. Therefore, I planted saplings in my village. After planting, it has been 10 years. Nowadays, fallen leaves and firewood are available. In my opinion, if all people plant 30 saplings each year, in the future Nepal will be good. Then, I think, it is not so difficult to get grass and firewood. Water and air

[42] 2(Southern dialect) + 318 (Classification number: Environmental Quality) + 01 (individual file number)

also will be good. People will also have long lives. If there are many trees, water will be good. Also it looks good to people. If there are trees, on the sunny day it is available to stay in the shade. If it rains, it will be used as an umbrella. The story of why we need trees, we have to speak to all the people. We, who are the knowing people, need to teach the unknowing one. We should not cut small saplings. If they cut, they need to cut big trees. If they cut one tree, they need to plant three saplings. We all need to think about this meaning. We need to plant saplings around houses and around streams.]

In the Sherpa text, we decided to change the Devanagari script to Roman script for easy parsing and the interlinear process. If, like Sherpa, the orthography is non-Roman, and is difficult to parse for the morpheme-breaks, one solution is to convert the whole text into Roman[43]. If it is in Roman, by using the computer parsing program, the gloss will be inserted automatically if the gloss of a morpheme has already been added to the glossary.

Each text will be divided into sentences, and have its own number of three digits. \tx is the main sentence of the text, and \mb shows the sentence parsed morpheme by morpheme, and \gl is the gloss of each morpheme. Finally, \ft is the free translation of the sentence. The title and first two sentences below are a sample from the text of 231801.

File name: 231801.txt

\ref 231801 001
\tx nepalla doŋbu gowu tamnye
\mb nepal-la doŋbu gowu tamnye
\gl Nepal-in tree need talking
\ft A story about the need of trees in Nepal

\ref 231801 002
\tx nepalla mi maŋšokki doŋbu cyetup tiraŋ misĩ
\mb nepal-la mi maŋšok-ki doŋbu cyet-up ti-raŋ misĩ
\gl Nepal-in man many-AG tree cut-Nml emph-only think-Con

\tx biruwa juwula sisaŋ nasam motoŋnok
\mb biruwa juwu-la si–saŋ nasam mo-toŋ-nok
\gl saplings plant-in who-Ag-also thought Neg-send-Fdj
\ft In Nepal many people only cut trees; nobody thinks of planting saplings.

[43] There are many programs to automatically convert computerized text from Devanagari to Roman script.

\ref 231801 003
\tx tikilagi ŋye ŋare yulla biruwa cuyin
\mb tiki-lagi ŋye ŋa-re yul-la biruwa cu-yin
\gl that-for I-Ag I-Gen village-in saplings plant-Pcj
\ft Therefore I planted saplings in my village.

From analyzing these texts, a glossary file is being built up. This glossary will later on be the basis of the dictionary. The following is a sample of the production after the analysis of text 231801.

\lx दोङ्बु
\ph doŋbu
\gl tree

\lx ताम्डे
\ph tamye
\gl story

\lx -ला
\ph -la
\gl in

\lx मी
\ph mi
\gl man

4.3 Word collection by semantic domain

4.3.1 The problem of corpus-based data collection in endangered languages

On the macro-structural level, word-frequency counts are very useful. Gouws and Prinsloo (2005) explain the advantages of this word-frequency count.

> A major advantage of such an approach is that, on the one hand, frequently used words will not accidentally be omitted and, on the other hand, that precious dictionary space will not be taken up by lemmas less likely to be consulted by the target user (Gouws and Prinsloo 2005:30).

However, there are problems in using the same method for data collection in endangered languages. The size of the corpus for endangered languages, if we consider the general situation of the budget for the development of the minority languages, could be around one million lexical items. If we were to exclude the lexical items that occur less than two times in frequency, it will

be good for controlling the quality of the collected data, but it would mean 42.5% would be excluded from the corpus of three million lexical items (Newell 1995:22). Furthermore, the words, which do not occur often in daily conversations, for example, specific names of such things as planets, trees, and materials for building a house, could easily be omitted and this applies especially to a certain group of words that are culturally prohibited to be used in spoken form, such as words referring to sexual organs or taboo-related words. If we consider a function of the dictionary as documentation of an endangered language, every word is important. As a lexicographer, if we lose it, it will be lost forever from the language. We should, therefore, include words occurring even just once or twice in the frequency count, if they are accurate and correct, because these words still belong to the language. In this case, frequency should not be the main criterion, but rather the role of documentation which is more important. In order to fill this gap resulting from low versus high frequency problems, two direct word-searching programs will be discussed in 4.3.2 and 4.3.3.

4.3.2 Newell's word-list elicitation approach

Newell's approach is a form of word-list elicitation through common and universally-known objects. Newell suggests that we use our imagination and cultural observations as follows.

> For example, one might think the word *stone* in Ifugao[44] would be elicited very simply. In fact, a little observation indicates that stones are a very important cultural item – stone wall building of rice terraces is just one area of complexity. We need to use our imagination, be inquisitive, pursue leads, formulate hypotheses, etc., when eliciting even what, on the surface, appear to be simple words of little cultural interest (Newell 1995:32).

For this approach, he advised using Murdock's Outline of Cultural Materials (1987), and he gave examples of suggested questions to elicit words.

- abaca: Where grown; how gathered; prepared by whom; uses.
- anus: Word-use taboos; term for taboo; list of tabooed words; under what circumstances such terms are/are not spoken; mixed company, brother/sister present.
- ashamed: Function of shame; cultural control; examples shame-inducing situations (Newell 1995:32).

[44] Ifugao is the name of a people group in the Philippines, where Newell had compiled the Batad Ifugao Dictionary.

4.3.3 Moe's semantic domain approach

Semantic domain is defined as a cluster of words in the mental network. Therefore, the words within the domain are linked by lexical relations, and the domains themselves are linked by lexical relations.[45] Saeed explains that a particular lexeme may have simultaneously a number of lexical relations, such as homonymy, polysemy, synonymy, opposites (antonymy), hyponymy[46], meronymy[47], member-collection[48], and portion-mass[49] (Saeed 1997:63-71).

Historically, there have been a few trials to utilize these semantic domains and lexical relations to elicit words for compiling a dictionary. To find new words, Beekman used four semantic domains, namely, objects, events, abstracts and relationals (Beekman 1975:365). Murdock approached it from the anthropological domains and made a list, which is in the *Outline of Cultural Materials* (Murdock, et al. 1987, Moe 2003:217). In *Roget's Thesaurus*, 1000 domains were developed. And Louw and Nida compiled their lexicon of the New Testament based on semantic domains, the *Greek-English Lexicon of the New Testament: Based on Semantic Domains*.[50] However, Moe's approach differs in two aspects from the previous approaches; one is semantic domains combined with lexical relations, and the other is the template approach (Moe 2003:217).

4.3.3.1 Semantic domain combined with lexical relations

Moe developed his semantic domains using nine main domains: the physical universe, person, language and thought, social behavior, home, work and occupation, physical actions, states, grammar and discourse. From these nine domains he established more than 1,800 sub-domains.[51] In each domain, he uses simple questions to elicit words which are related lexically. Sample questions are as follows:

---

[45] Quoted from: http://semdom.org/description [2016: June 30]
[46] Hyponymy means a more general/generic word, e.g., the hyponymy of dog and cat is animal (Saeed 1997: 68).
[47] Meronymy is a part-whole relationship. Wheel, engine, and door are parts of a car (Saeed 1997: 70).
[48] Member-collection is a relationship between the word for a unit and the word for a collection of the units. A fleet is a collection of ships (Saeed 1997:71).
[49] Portion-mass is the relation between a mass noun and the usual unit of measurement or division, e.g. a sheet of paper (Saeed 1997:71).
[50] Louw and Nida explained their principal reason for compiling the Greek-English Lexicon based on semantic domains was the inadequacy of most existing lexicons (Louw & Nida 1988: viii).
[51] For further details of Moe's semantic domains go to: www.semdom.org

What words refer to singing? sing, serenade, warble, yodel, burst into song
What words refer to singing without using words? hum, whistle
What words refer to a person who sings? singer, vocalist, soloist
What words refer to a group of people singing together? choir, chorale, singing group, duet, trio, ensemble
What words refer to something that is sung? song, singing, tune, melody
What types of songs are there? lullaby, hymn, psalm, carol, national anthem, lament, ballad
What words refer to a part of a song? verse, chorus, theme, note, melody, harmony
What words describe how well a person sings? beautiful singing voice, can't carry a tune in a bucket, sing on/off key, monotone
What words describe how high or low a person sings? pitch, soprano, alto, baritone, bass
What words describe whether or not people are singing the same thing together? sing in unison, sing in harmony, sing the melody/harmony

4.3.3.2 Semantic Domain by template[52]

Moe raised two problems in the world of lexicography. The first is that, for the lexicographers, the task to complete dictionaries for the whole world is so huge. In the world, there are perhaps 6,000 languages, and if we guess that there are at least 20,000 words in each language, the total words, which we need to collect and describe, are 120,000,000. This is the place where we have to think about the efficiency of the work. The second problem is that the funds to solve this problem are so limited. Especially the funds for minority languages are really limited and, even more so, for the endangered languages. These two problems are the background for the development of a template. A template is a good tool for lexicographers to collect information about the words faster and more systematically. Moe produced a universal template, which is based on cross-linguistic research, and on features which the lexicographers could encounter in each domain. In each domain, three things were provided: (1) a simple statement of the central idea of the domain, (2) elicitation questions that would prompt a person to think of words that might belong to the domain, and (3) sample words from English (Moe 2001:150, Moe 2003:216-220). Moe shows one sample (2.4.1[53]) of the template from the 1,800 domains as follows:

2.4.1 See
What words refer to seeing something (in general or without conscious choice)?
see, behold, come into view
What words refer to consciously looking at something?

[52] This template called the Dictionary Development Program by Moe, was mentioned in 2.2.3.
[53] 2.4.1 is a serial number of Index of Semantic Domains.

look at, view, observe, scan
What words refer to looking at something in order to learn?
watch, scrutinize
What words are used for looking at something for a long time or in amazement?
stare, gaze, gape, gawk
What words are used for looking at something for a short time?
glance, cursory glance, look at briefly, (eyes) flicker over
What words refer to the sense of sight?
sight, sense of sight, vision
What words refer to someone who sees?
observer, beholder, witness
What words refer to a group of people who are watching something?
audience
What words refer to what is seen?
sight, view (Moe 2003:218)

Moe reported the result of the tests after he used this template to collect words from three minority languages as in Table 10.

Language Name	Date	Length of test	Participants	Collected words
Lugwere	May 2001	10 days	15 people	10,000 words
Lunyole	Jan. 2002	10 days	30 people	17,000 words
Kitharaka	Feb. 2002	8 days	12 people	12,000 words

Table 10. Test results of Moe's template (Moe 2003:218)

We, as the Sherpa dictionary team, used this template on 7th April of 2005 for two weeks with 6 Sherpa people, and collected 6,500 words[54]. After deleting all the duplications, Nepali loan words and phrase type of words, we got about 5,500 lemmas. Since we already had 4,000 lemmas through text collections, we found that the new lemmas were more than 1,500. Some examples from this template of new lemmas that were added to our dictionary file are:

\lx टेताङ
\de early in the morning
\is 1.1.1[55]
\sd Sun

[54] The reason why the number of collected words was smaller than the numbers in the cases of Moe, was that we already had collected 4,000 lemmas through text collections and we could delete the duplications and phrase type of words. Another advantage was that our team was trained in the Sherpa spelling system, so there weren't cases of the same word spelled in different ways.
[55] \is: index of semantics, which is in Moe's template.

\lx ड्ियमी शारीन
\de sunny
\is 1.1.1
\sd Sun

\lx हल्गी
\de galaxy
\is 1.1.1.2
\sd Star

\lx कन्टुक्पा
\de six stars of constellation
\is 1.1.1.2
\sd Star

\lx हलीथोड्बा
\de three stars of constellation
\is 1.1.1.2
\sd Star

We could also see possible challenges in the process of collecting words by semantic domains with a template. I personally managed one workshop for the Sherpa project and participated as an observer in the Tharu[56] workshop. In both workshops, I found two similar problems; the one was difficulties in handling the lexical relations. Since most of the people involved in the dictionary work were not educated people, they were not familiar with the generic-specific distinction. Ed Boehm reports this in his feedback after his workshop for the Tharu project:[57]

> The biggest problem is the participants not understanding the generic-specific distinction. In training it would be good to take them through several sections that go from generic to specific in several different domains. It may also be possible to include a hint in the more generic domain that we are only looking for a few words that summarize the whole section. It may be better to work from specific to generic, but that will also take some training in how to distinguish a generic from a specific area. [58]

[56] Tharu is a language of Nepal belonging to the Indo-Aryan language family.
[57] The Tharu Word Collection Workshop was held March 22-31, 2005 in Nepalganj, Nepal. The leader was Ed Boehm.
[58] This report was a sent to me personally by Boehm after the workshop.
URL:http://www.sil.org/resources/archives/66562 [2016, August 4].

The other was the variants in dialects even in the same dialect speaking areas. This dialect problem is actually never-ending in the process of all dictionary projects. Probably, after considerable struggles with dialects, finally a dictionary could be produced for a standardized variety. This, I believe, is a major role for which a dictionary exists.

4.4 Chapter summary

After considering the conceptualization of the dictionary compiling, we have to move to the data collection. This data collection should be on the basis of the genuine purpose of the dictionary to be compiled. Personally, I still continue to struggle with the dialects and variants of the Sherpa language, because this language has not been standardized yet. So, I had a special focus on the study of this language in 4.1 as a prerequisite to the data collection process. Then, two complimentary methods were explained, i.e. corpus-based data collection and semantic domain-based data collection.

As a lexicographer, it is a privilege to find the hidden lexical items from the jungle of information, and present them as lemmata in a dictionary. It is important to find each and every word that may possibly otherwise disappear from the language, and include them in the dictionary. Documenting these endangered words is a part of the responsibility of the lexicographer. And also it is the responsibility of a lexicographer to collect qualified data in a proper way without any bias of religion, regionalism or history. We are the people who are reconstructing a dying language.

Chapter 5. The structures of a Sherpa Dictionary

5.0 Introduction

In the previous chapters, we focused on general lexicographical theories and the methods of collecting the lemma candidates. Finally, it is time to discuss in more detail the envisaged Sherpa Dictionary. This chapter will contain the core of the model of this dictionary with a specific focus on the different structures to be employed. First, as meta-lexicographical criteria, the typological and functional models of a Sherpa Dictionary will be explained. Then, the frame structure, macrostructure and microstructure of the Sherpa Dictionary will be described.

5.1 Meta-lexicographical criteria of a Sherpa Dictionary

Gouws and Prinsloo (2005) explained three criteria to improve the quality of dictionaries, which are also applicable to bilingual dictionaries, introduced by Kromann et al. (1991:2713). They are the user aspect, the linguistic aspect and the empirical aspect. The user aspect is mainly directed at the target user group and their lexicographical needs. The empirical aspect shows the establishment of relevant databases. And finally, the linguistic aspect treats the relations between the lexical fields of the source and the target language (Gouws 1996:18). Since the empirical aspect has already been discussed in the Chapter 4, the user aspect will be covered in 5.1, and the linguistic aspect in the rest of Chapter 5.

5.1.1 Typological model

Since the general typology has already been mentioned in 3.2, in this section, the typological model of the Sherpa Dictionary will be discussed with more focus on features that should be considered for inclusion. Gouws and Prinsloo identified the need for dictionaries in South African speech communities as urgent by saying that:

> The typical needs of the members of the South African speech communities demand the ‘speedy’ availability of dictionaries (Gouws and Prinsloo 2005:45).

Moe also mentioned that 120,000,000 words need to be collected for 6,000 languages (Moe 2001:150). The urgency and the large number of unfinished dictionaries demand that lexicographers have to be very wise and they have to make realistic choices in the compilation of dictionaries (Gouws and Prinsloo 2005:45). In this regard we have to pay attention to what

Hausman says about the lack of harmony between lexicography and the general public. He said this lack of harmony results from a conflict between a *dictionary culture* and *user-friendliness* in lexicography (Hausman 1989:13). As lexicographers, we should consider the situation of the target language and the reference skills of target users to make the right typological choice and decide the functions of any dictionary. Since we have already debated the theoretical typology of the envisaged dictionary in 3.1.2 and 3.2.3, in the next sections of this chapter, we will only discuss the specific typology based on the situation of the target Sherpa dictionary users. Here, I will refer to each characteristic as a feature not a type.

5.1.1.1 The issue of script: The feature of multi-script

The script issue was already discussed in 2.2.3. In summary, therefore, within the Sherpa society, there is a strong conflict with regard to the script issue. Ideally, Sherpa people prefer to use the Tibetan script, because this script is strongly connected to their Tibetan identity. However, practically, literate people, who are familiar with this script, are less than 10 % of the whole Sherpa population, and these are mostly related to the Tibetan religion, such as monks and nuns. Most of the Sherpa people are more exposed to the Devanagari script, which is the national script in Nepal. If we take this conflict among the potential users into consideration, the main lemmata would be presented in two scripts, the Devanagari script and the Tibetan script respectively before the item giving the pronunciation is presented. This will be done as follows, the first word is the lemma in Devanagari, and the next one is the same but in Tibetan script[59]. This is followed by the item giving the pronunciation.

पेमा, དེ་མ། [pema[11]] *n*. sand, sandy soil. बालुवा, बलौटे माटो. WD: पेप्शोक

In this case, the sorting order of the main lemma in the macrostructure will be done according to the Devanagari forms. In the printed dictionary the Devanagari lemmata will be the principal guiding elements of the dictionary articles. In the Internet dictionary both alphabetical systems will be able to be used.

[59] From the publisher's point of view this will be very complicated, because the size of the Tibetan font needs to be bigger than 14, and the line space will be 1.5.

5.1.1.2 The issue of Sherpa language proficiency: The feature of multilingualism for Nepali

In the Sherpa Sociolinguistic Survey Lee (2003:89) tested language use by asking people what language they use in different places. For the question concerning 'at home', we could see their language use within their family. For 'in the temple' we could guess their religious lives, and for 'at the market' their social lives in the mixed culture with the Nepali language. The result was as follows:

Question	**Response**					**Sample size**
What language do you use….	Sherpa	Tibetan	Nepali	Sh /Nep	Sh /Tib	
at home	82%	-	14%	4%	-	50
with friends	40%	-	20%	40%	-	50
in the village	58%	-	12%	30%	-	50
in the temple	82%	6%	4%	2%	6%	50
at the market	6%	-	66%	28%	-	50

Table 11. Survey result of Language Use (Lee 2003:89)

Even though the data are not very recent, as we do not have a recent survey, with these data, we can still judge the general language use of the target users of the Sherpa Dictionary. We can assume that there will be five possible groups of language use among the Sherpa people. Among the five groups, we can analyze that there will be three types of target users.

	In temple	At home	With friends	In village	At market	Target users
1	Sherpa	Sherpa	Sherpa	Sherpa	Sherpa	Sherpa mother tongue speaker
2	Sherpa	Sherpa	Sherpa	Sherpa	Nepali	
3	Sherpa	Sherpa	Sherpa	Nepali	Nepali	
4	Nepali	Sherpa	Nepali	Nepali	Nepali	Minimal Sherpa speaker
5.	Nepali	Nepali	Nepali	Nepali	Nepali	Nepali speaker

Table 12. Five possible groups of language use among Sherpa people

They are 1) Sherpa mother-tongue speaker, 2) minimal Sherpa speaker who uses Sherpa language only at home, 3) Nepali speaker whose mother tongue is Nepali. This analysis shows that the Sherpa Dictionary should be multilingual as in 3.1.2.4. One of the meta-languages should be Nepali for the Sherpa people whose mother tongue is Nepali, and for the minimal Sherpa speakers.

5.1.1.3 The issue of internationalization: The feature of multilingualism for English

There are three reasons to have a feature of multilingualism for English. First, there are no official statistics known about Sherpa people who are living abroad, but the non-official numbers of Sherpa residents in New York alone are known to be more than 2,500[60]. This information leads us to guess that about ten thousand Sherpa people live outside of Nepal. Historically, Tenzing Norgay Sherpa was in India when he was selected as a guide by Sir Edmund Hillary for his 1953 Mt. Everest expedition team. In comparison with other language groups of Nepal, Sherpa people had the most contact with foreign trekkers, so more Sherpa people are probably living abroad. Most of these people and their succeeding generations are not able to speak the Sherpa language. For this target group of Sherpa people, the Sherpa Dictionary is necessary to be multilingual with English as one of the languages.

Secondly, in the education system of Nepal, a few schools use English as a medium of instruction rather than Nepali. This change of language policy is more common in the capital city, Kathmandu.

The final reason for the dictionary to be multilingual with English as a language is the openness of the Sherpa language to international readers and scholars, who want to learn or research this language.

5.1.1.4 The issue of dialects: The feature of standardization

The Sherpa Dictionary will be overall-descriptive, which means it will include dialects. The dialects of Sherpa have already been discussed in 2.2.2. There are three dialects, and among these the Southern variety is the prestigious dialect. The Southern dialect will be used as the source language form from which the primary lemma candidates will be selected; whereas the user will be cross-referred to the variants from the other two dialects. For example, the main lemma (पेमा) has a full description with a cross-reference to पेप्शोक, the variant from the Western dialect (WD). This variant is not the guiding element of a default article but only of a

[60] Source: https://en.wikipedia.org/wiki/Sherpa_people#Mountaineering

cross-reference article which guides the user to the lemma (पेमा) with its comprehensive treatment. This is illustrated in the following articles:

पेमा, ཞ་མ། [pema[11]] *n*. sand, sandy soil. बालुवा, बलौटे माटो. WD: पेप्शोक
पेप्शोक *See* पेमा

An overall-descriptive dictionary is, in one sense, a repository of all dialects, but, in another sense, the making of such a dictionary is primarily a process to create a standard Sherpa dictionary as mentioned in 3.2.3. For endangered languages, the compilation of a dictionary also is the execution of a process of standardization that elevates one variety to the standard.

5.1.1.5 The issue of website uploading: The feature of a website dictionary

The compilation of a printed dictionary is limited by the budget of the publisher with regard to its size, which affects the structures, the extent, and the contents of its articles, etc. This is the reason why the compilers have to consider the possibility of an internet dictionary or a dictionary on a website. When they start to compile a dictionary, they should investigate a computer program which has a system to upload to a website. For example, the Nepali-Korean dictionary (Lee 1999), which was produced by Toolbox, could easily be uploaded to the website, and it is now linked on www.naver.com[61]. The advantage of this website dictionary is that it is not limited by the extensiveness of the data. And also all sorting systems are available for the ordering of different languages, which makes the multilingual dictionary much easier for the users. So, the lexicographers, who want to compile a dictionary of the endangered languages, should keep in mind this issue of website uploading from the early stage of the dictionary conceptualization and have to apply it in all processes of the dictionary compilation.

5.1.1.6 Typological models of the envisaged Sherpa Dictionary

In conclusion, the envisaged Sherpa dictionary will have a hybrid typology, which has the following typological features:

- Linguistic dictionary
- General dictionary

[61] Cf. http://nedic.naver.com/

- Synchronic dictionary
- Dictionary with multiscript lemma in Devanangari and Tibetan[62]
- Multidialectal dictionary
- Standard dictionary based on the Southern dialect
- Dictionary based on spoken language
- Dictionary based on corpus and semantic domain
- Dictionary uploadable to a website

5.1.2 Functional model

The Sherpa Dictionary will have different features.

a. *A language documentation feature*

Sherpa, as an endangered language, needs to be documented for its preservation. All lemmata, dialectical differences and variations should be collected and included to the best of the compiler's ability. Exact pronunciations with tone markers, if the language has them, definitions with illustrations, and cross-references are crucial for the language maintenance.

b. *A didactic feature*

In school, this Sherpa Dictionary will be used as a teaching tool to teach the Sherpa language to the young students. It will help with the following questions:

- What is the standard spelling?
- What is the standard dialect?
- What is the difference between the standard dialect and other dialects?

The Sherpa language usage given in the illustrations will be a good resource to learn the language for Sherpa people whose mother tongue is not Sherpa.

In adhering to the function theory this Sherpa Dictionary will have both a communicative and a cognitive function.

A communicative function

This Sherpa Dictionary will be used for the following situations (Tarp 2008:53, cf. 3.1.1.3):

- Production of text in the Sherpa language
- Reception of text in the Sherpa language
- Production of text in Nepali and English.

[62] This is actually a double-headed lemma sign consisting of two lemmata that each function as a guiding element.

- Translation of text from Sherpa into Nepali and English
- Proofreading or marking of texts produced in Sherpa
- Proofreading or marking of texts translated from Sherpa into Nepali and English

A cognitive feature

This Sherpa Dictionary will be used for acquiring the following types of knowledge (Tarp 2008:119-120, Cf. 3.1.1.3).

- Knowledge about the Sherpa language
- Knowledge about the Sherpa vocabulary
- Knowledge about the Sherpa grammar
- Knowledge about the Sherpa culture

5.2 A structural description of a Sherpa Dictionary

Like all utility products, a language dictionary has its own genuine purpose, which has already been explained in 3.3.3 and 3.4.1. The genuine purpose of a dictionary, according to Wiegand (1998), is to assist the target user who uses the dictionary in a typical usage context to achieve a successful dictionary consultation procedure by reaching the goals that motivated the search. Dictionary compilers for endangered languages should consider this genuine purpose more specifically, because the language documentation for endangered languages is still under development for the language documentation. For some languages, such a dictionary could be the first language documentation in the history of that language. The users may not be familiar with the use of linguistic references and may struggle to understand the dictionary's structure, to read the orthography, and to identify the alphabetical ordering, etc. It is the responsibility of the lexicographers, who understand these situations of users, to plan and compile a dictionary that gives easy access to the lexicographic data. For the user, excellent data that are hidden in a corner, are completely useless. This process of planning and compiling an easy-to-use dictionary should start thinking and planning for these things in the early stages of dictionary conceptualization, so that, from the beginning of the project, data could be collected in accordance with the type, function and structures of the envisaged dictionary.

Recent discussions concerning meta-lexicography suggested that dictionaries should be regarded as carriers of text types (Wiegand 1996). This means that a dictionary is a "big" text made up of different kinds of texts (Gouws and Prinsloo 2005:57). Within the book structure of a dictionary there are three main areas, i.e. the front matter, the central list, and the back matter. The central

list is the main and compulsory part of the dictionary. The front matter section is located before the central list, whereas the back matter section follows the central list. In terms of the structure of dictionaries there are two approaches, i.e. a word book structure and a word list structure. The word list structure focuses only on the central list, but the word book structure is focused on the front and back matter sections. The texts included in the front and back matter sections are called *outer texts* (Gouws and Prinsloo 2005:57). Outer texts are not compulsory, but the front matter section can be regarded as being used more often, because it should include a text presenting the user guidelines. Where the central list of a dictionary is complemented by both front and back matter sections the dictionary displays a *frame structure* (Kammerer and Wiegand 1998, Gouws and Prinsloo 2005:57). In the following sections aspects of this frame structure of the envisaged dictionary will be explained. The discussion will first focus on the outer texts, and then the central list will be discussed.

5.2.1 Outer texts of a Sherpa Dictionary

Outer texts are very important in a dictionary for an endangered language for at least three reasons. First is the understanding of the language itself. If the language does not have any grammatical analysis, the dictionary has an additional role, and the outer texts could provide an outline of the grammar of the language. This grammatical study will not only be good for the mother-tongue Sherpa speakers in helping them to understand their language, but also for Nepalese and foreign scholars to be able to grasp the language. Second is the alphabetical order. Since the issue of alphabetical order is making its first appearance in the language's history, nobody would be able to identify the specific alphabetical order, because there are a few newly combined alphabetical characters from the Devanagari system used in the Sherpa orthography. For the users, this order will be a continual problem to remember in consulting the dictionary. Third is specific information about the language, i.e. the lists of numbers, colors, irregular verbs, etc. The compiler has a freedom to either include outer texts in both the front and back matter section or to put them only in the front matter section. For the envisaged Sherpa Dictionary, both these venues for outer texts will be employed. Fourth is the understanding of Sherpa culture, i.e. the religious terms, wedding related costumes, women's ornaments, domestic and agricultural materials, etc. On the basis of this information, the outer texts will have a cognitive function

helping the users to obtain knowledge about the Sherpa language, its orthography, grammar, and Sherpa culture (cf. 5.1.2).

Kammerer and Wiegand (1998) made a distinction between integrated and non-integrated outer texts. When the outer texts share the feature of presenting data regarding the subject matter of the dictionary in order to accomplish the genuine purpose of the dictionary, with the central list of that dictionary, it is called integrated outer texts. Non-integrated outer texts function alongside the central list and are not needed to retrieve information presented in the articles of the central list or to contain data relevant to achieving the genuine purpose of the dictionary (Gouws and Prinsloo 2005:58-59). If we make a distinction in this regard between integrated and non-integrated outer texts, the Sherpa Dictionary will have integrated outer texts, so that outer texts will share the feature of presenting data with the central list of the Sherpa Dictionary that will focus on the subject matter of the dictionary and help to achieve its genuine purpose.

With regard to the data distribution structure, Bergenholtz, Tarp and Wiegand (1999:1779) identified two main types of data distribution structure, i.e. a simple data distribution structure and an extended data distribution structure. A simple data distribution structure is found where the central list is the only venue for the data distribution; whereas an extended data distribution structure is found where outer texts or parts of outer texts are employed to accommodate data as part of the procedure of data distribution (Gouws and Prinsloo 2005:58). In this regard the Sherpa Dictionary will have an extended data distribution structure, because both front matter and back matter texts are employed as part of the procedure of the data distribution.
In the following sections the specific contents of outer texts will be explained.

5.2.1.1 The front matter texts

5.2.1.1.1 Title page

On the title page, there will be the title of the dictionary in both the source language and the target language(s) of the dictionary, the name(s) of the authors or editors, and the publisher with the location and the year of the publication. Optionally, name(s) of institutions or universities, that sponsored the dictionary project or under whose auspices it was implemented could also be included.

5.2.1.1.2 Recommendation letters

Recommendation letters will be located after the title page, showing that this dictionary is authenticated by a person or by the society of the language. The author(s) of the recommendation letters will usually be political leaders or representatives of the society or well-known scholars. For endangered languages, there is no official approval process for the dictionary, so this is a kind of recognition letter of an important person, who has respect and honor within the society. Landau (1984) says that this kind of comment can give authority to the dictionary.

> Thus, front-matter articles are often written by prominent scholars or educators in an attempt to establish the authority of the work and lend it prestige (Landau 1984:148).

If the dictionary project is sponsored by an outside funder, a recommendation from this financial supporter is also possible. This is not just a recommendation, but also an approval of the whole dictionary project by the funding agency.

5.2.1.1.3 Table of contents

The table of contents in a dictionary is a map of all components of the dictionary showing where they are located. Gouws and Prinsloo (2005) reminded us that the dictionary is a "big text", and explained the purpose of the table of contents as follows:

> The purpose of a table of contents should not only be to give an overview of the contents of the dictionary but also to increase the access of the dictionary as a big text by means of an indication of page numbers ensuring a rapid progress to the different texts constituting the big text. In this regard, the table of contents puts the user on the dictionary internal search route (Gouws and Prinsloo 2005:166-167).

5.2.1.1.4 Preface

The preface follows the table of contents, and can be replaced by an introduction or by a letter from the editors. The components of the preface could include:

- The history of the dictionary project
- The purpose of the dictionary
- Introduction of editors or board of editors
- Introduction of the financial supporters, if any
- Some highlights of the dictionary process
- Thanks and appreciation to the people who were involved.

5.2.1.1.5 User's guidelines

Many researchers found that the average dictionary user does not read the user's guidelines text before they consult a dictionary (Gouws and Prinsloo 2005:85). However, this guideline text is the best place to show the target user how to understand the structure of the dictionary, indicate the different types of data in the dictionary and the venues of the data and to show them how to retrieve the required information from the lexicographical data being offered. The dictionary compiler should use this text to give a comprehensive explanation to inform even users with no experience of dictionary use how to use the specific dictionary in an optimal way. In this regard, these guidelines will include an explanation of the macro- and microstructure of the dictionary. Some possible topics of these texts are as follows:

- The target readers
- The main dialect, and the treatment of different dialects with regard to the selection of lemmata and cross-reference entries
- The treatment of the variations within the same dialect
- Alphabetical order and the location of consonant clusters, diphthongs, and special characters
- Understanding the pronunciation, and tone markers
- Differentiating homonyms
- 'The parts of speech' included in this dictionary. A complete list with their abbreviations should be given[63].
- The special mark for loan words, and how the language of origin is indicated.
- Cross-references. A list of all kinds and their abbreviations should be given.
- An explanation of the use of structural indicators and a list of structural indicators, which are used within the article structure[64].

5.2.1.1.6 A phonological and grammatical summary

Preparing a phonological and grammatical summary is not an easy job for the dictionary compiler, but this summary is crucial data for endangered languages, if these languages do not have an official grammatical analysis yet. Therefore, this phonological and grammatical summary will provide a clear window to display the structures of the language, not only for the mother-tongue speakers, but also for other users and foreign scholars. Possible components of such a phonological and grammatical summary are as follows:

[63] These abbreviations could be gathered together and located in a section of Abbreviations.
[64] cf. 5.2.4.2

- Introduction to the language
- Phonological summary
  - Vowel phones and vowel chart
  - Consonant phones and consonant chart
  - Tone summary, if they have tone system
  - Consonant clusters
  - Diphthongs
  - Phonological differences in the different dialects
- Morphological aspects
  - Morphology of noun phrases
  - Morphology of verb phrases
- Clause and sentence structure

5.2.1.2 The back matter texts

Gouws and Prinsloo (2005) enforce the need for the back matter texts by saying,

> The use of back matter texts which contain lists of items that also feature as lemmata in the central list of the dictionary necessarily elevates the dictionary to a poly-accessible source because there is more than one position from where a user can find access to a specific lemma (Gouws and Prinsloo 2005:62).

The components could be 1) grammatical components, 2) cultural lists, and 3) items which are used in daily living.

5.2.1.2.1 Grammatical components

In this section, we should not overlap the contents with the mini grammar summary in the front matter section, but include any grammatical words, which need to be arranged within a group to clearly see the grammatical differences, would be good to put in this section. One example will be verb conjugations especially for irregular verbs, i.e. the present, past conjunct, past disjunct, imperative, and nominalized forms. If possible, all irregular verbs could be gathered in this list, so that non-mother-tongue speakers, who search for these complicated verb conjugations, can visit this section.

Each language has its own irregular conjugations of verbs and nouns. In the Sherpa language, the genitive marker (Gen) is *–ki*, but there are many irregular markers for different word types. Examples are:

yulki
yul-ki
village-Gen

doŋbi
doŋbu-i
tree-Gen

ŋye
ŋa-e
I-Gen

Numerals are also a good category for this section. Both cardinal and ordinal numbers should be explained together. For the Sherpa cardinal number system the following items have to be shown:

- 1 – 20
- 30, 40, 50, 60, 70, 80, 90, 100
- 1000

For the Sherpa ordinal numbers, from the *first* to the *tenth* need to be shown as follows:

ताङ्बो (first), ङिवा (second), सुम्बा (third), ज्यिवा (fourth), etc.

5.2.1.2.2 Cultural components

Cultural components in Sherpa would be such things as religious objects, wedding garments, Sherpa women's jewelry, Sherpa domestic materials, agricultural items, and village names along the trekking routes from Jiri to the Base Camp of Mt. Everest. Examples are:

- Religious objects with a photo: names of specific parts of the Tibetan Buddhist shrine, flags on the pole where the Buddhist scripture are printed, flags on the shrine and on top of houses in the village, etc.

- Wedding dress and names of bride and groom's dress

- Sherpa women's ornaments

- Sherpa domestic materials: cooking pot, dish, plate, cup, pan, knife, etc.
- Agricultural items: axe, hoe, threshing instrument, etc.
- Village names along trekking routes from Jiri to the Base Camp of Mt. Everest. This will be a useful guideline for the tourist who will trek in this area.

5.2.1.2.3 Items from daily life

The following lists could be included as back matter texts:

- Names of the colors
- Names of flowers growing in the Himalayas
- Names of birds inhabiting the Himalayas
- Traditional names of Sherpa villages, which are different from the official names
- Kinship terms

5.2.1.2.4 References

References consist of two components: general books, which were used as references for the study of the language, the people and their cultures, and dictionaries for the structure of dictionary.

5.2.2 The central list of a Sherpa Dictionary

Gouws and Prinsloo (2005:62) said that the central list is the most salient component of a dictionary displayed within the frame structure. To make optimal use of a dictionary, a lexicographer should plan the data distribution structure early in the dictionary conceptualization process. This data distribution structure will determine all structures not only for the outer texts, but also for the central list. The envisaged Sherpa Dictionary will be a trilingual dictionary with Sherpa, English, and Nepali. It will be mono-directional, from Sherpa to English and from Sherpa to Nepali. There will be only a single macrostructure in the central list. The following structures of the central list will be discussed in the subsequent sections:

- Macrostructure
- Microstructure

5.2.3 Macrostructure of a Sherpa Dictionary

Hausmann et al (1989:208) simplified the relationship between macro- and microstructure as in the diagram below. Let's say there are lemmata X, Y, and Z, and also there are data[65] presented as part of the treatment of X, Y, and Z. The ordered set of all the lemmata of the dictionary forms the macrostructure (the vertical rectangle in the diagram). The lemma and the whole set of data items, which are related to the lemma, form the dictionary article (the horizontal rectangle in the diagram). The ordering structure of data within the article is called the microstructure (the oval in the diagram). In the classical conception of microstructure, the lemma does not belong to the microstructure.

[65] As a result of recent discussions among lexicographers, the terminology for the 'items of information' mentioned in Hausmann has been changed to data. According to that discussion, data are objective; whereas information is subjective depending on the reference skill of the readers. So, data are a more correct term than information (from personal discussion between Lee and Gouws on 11th October 2016 at Stellenbosch University).

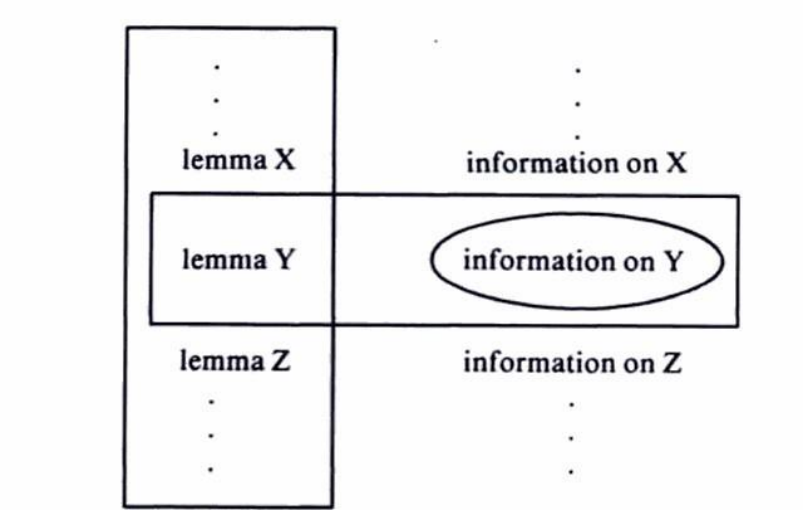


Figure 10. Simplified visualization of macro- and microstructure of the dictionary (Hausmann et al 1989:208)

In the following subsections the macrostructure of the Sherpa Dictionary, i.e. the lemmatization strategies, different types of lemmata and articles, and the ordering of lemmata will be explained.

5.2.3.1 Lemmatization strategies

A lemma is the basic particle or word(s), which could be the head-word or guiding element in the dictionary article. Especially in the printed dictionary, there is always a limitation of space, so we need a lemmatization strategy for the selection and presentation of lemmata. According to Gouws and Prinsloo (2005:67):

> Lemmatisation can be defined in an over-simplified way as the selection of a specific form from a given paradigm to be used in a dictionary as the starting point for information retrieval.

For the envisaged Sherpa Dictionary there will be three lemmatization strategies.

5.2.3.1.1 Frequency-based strategy

The Sherpa Dictionary will get the lemmata both from the corpus and the different semantic domains. The choice of lemmata is often determined by the frequency of their use in the corpus. Theoretically, we would select the high frequency lemmata as first priority for inclusion, putting the high frequency indicator before the pronunciation, and also mark the middle frequency and the low frequency lemmata as follows:

❸: high frequency[66]
❷: middle frequency
❶: low frequency

And an example article is:

पेमा, པེ་མ། ❷ [pema[11]] *n*. sand, sandy soil. बालुवा, बलौटे माटो. WD: पेप्शोक.

However, in the case of endangered languages, we should not limit the lemma selection of the dictionary to high frequency words, as stated in 4.3.1. If we consider the Sherpa language as an endangered language, the frequency should not be the main criterion for the selection of lemmata, because the role of documentation is more important in this dictionary. So, we will include words occurring even just once or twice in the frequency count, although it is still very informative to have the frequency of usage noted in the dictionary as above.

5.2.3.1.2 Lemmatizing verbs

Like other Tibeto-Burman language groups, the verb system of the Sherpa language is quite complicated. For the lemmatizing, we have to choose what will be the head word for the verb paradigm. The table below shows the Sherpa verb variations in different paradigms.[67]

Change type	**Stem**	**Present**	**Past Conjunct68**	**Past Disjunct**	**Impera -tive**	**Nominal-ized form**	**Meaning**
regular	pul[2]	pul-giwi	pul-in	pul- suŋ	pul	pulup[22]	to push
stem final cons. added	pe(t)[2]	pe-kiwi	pe**t**-in	pe-suŋ	pe	petup[22]	to choose
stem final cons. dropped	kə(k)[2]	kə-**k**iwi	kə-yin	ka- suŋ	ko	kəkup[22]	to break
vowel ə>a>o	jyəl[1]	jyəl-giwi	jyəl-in	jyal-suŋ	jyol	jyəlup[11]	to visit (Hon.)
vowel i>a>o	liŋ[1]	liŋ-giwi	la-yi	la- suŋ	lo	liŋgup[11]	to take
vowel o>a>o	toŋ[2]	toŋ-giwi	toŋ-in	taŋ- suŋ	toŋ	toŋgup[22]	to send
cons. vd>vl & tense lo>hi	ɖol[1]	ɖol-giwi[111]	ʈol-in[22]	ʈol- suŋ[22]	ʈol[2]	ɖolup[11]	to untie

[66] The reason why I do not use the exact number of frequency here is that I do not have all of the frequency data yet. At this moment I just call them as high, middle, and low frequency.
[67] This table does not represent all verb variations, but just examples to decide verb lemmatization.
[68] Kelly (2003:358) explained the conjunct/disjunct relationship. The conjunct form occurs with first-person actors in main clause statements and as second-person actors in interrogatives. The disjunct occurs with the second- and third-person actors in declaratives and first-person actors in interrogatives.

vowel e>a>e & cons. vd>vl	gyek[1]	gyek-iwi	kyek-in	kyak-suŋ	gyek	gyekup[11]	to block
vowel e>a>o & cons. vd>vl	ɖel[2]	ɖel-giwi	ʈəl-in	ʈal- suŋ	ʈol	ɖelup[22]	to separate
combined two different stems	ɖo[2] ter[2]	ɖi-wi ter-kiwi[222]	gəl-in bin-in[11]	gal- suŋ bin- suŋ[11]	gyuk bin[1]	ɖop[2] terup[22]	to go to give

Table 13. Sherpa verb variations

Here, we have to decide two things. First, which form among these paradigms will be the lemma? Second, shall we include these paradigms in the verb articles? If included, where? I already mentioned that the whole list of irregular verb forms will be presented in a back matter text, so here we need to discuss about the macrostructure of this dictionary.

5.2.3.1.2.1 Lemmatizing verb: Stem vs. word

Prinsloo (2011:173-185) clearly explained the dispute with regard to the lemmatizing of verbs whether as a stem or a word. Both a stem and a word as a lemma have their advantages and disadvantages. As African language verbs *freely* and *productively* combine with a huge number of prefixes, a group of verbs, of which the stem or root are the same, will be scattered throughout the dictionary because of the alphabetical order. This will be a negative aspect for the first learner. However, identifying the right stem from the verb, which has variations, would also be a big challenge for the new reader (Prinsloo 2011, Gouws and Prinsloo 2005:78-81). For the Sherpa language there are two possible candidates as a lemma, one is to use the stem or nominalized form, which is a combination of a stem and a nominalizer (-up)(cf. Table 11). Actually the nominalized form was used as a lemma in the early stage of the Sherpa Dictionary. If the verb follows a regular verb paradigm, the nominalized form works well. However, if the verb is irregular, the vowel and consonant changes in the verb will not be shown in the nominalized form, even though this problem is almost the same in the case of the stem which is the second possible candidate. The main issue is which one is easier for the new learner, and also more economical as far as space concerned. After taking into consideration these advantages and disadvantages, I decided to use the stem as a lemma for verbs in the Sherpa Dictionary. Therefore, the rule of the verb lemmatization is:

1) The verb stem will be the lemma.

2) In the case of regular verbs, the verb conjugations will not be shown in the article.

3) In the case of irregular verbs, the verb conjugations will be shown under the stem lemma.

4) In the case of irregular verbs, the nominalized form will also be given as a lemma but without a detailed article, and only with a cross-reference to the lemma representing the main verb. All the other verb variations, e.g. present (pr), past conjunct (pc), past disjunct (pd), imperative (im), nominalized form (nom), should also be included as main lemmata, but still it will present the problem of the size of the dictionary from the perspective of the publisher. This would not be the best solution for all irregular verbs, so continued study regarding this is needed. The examples are in the next section, 5.2.3.1.2.2 after the explanation of the location of verb variations.

5.2.3.1.2.2 The location of verb variations

The variations of irregular verbs will be given in two places, i.e. as a back matter text and as an entry within the individual article between the pronunciation and the part of speech. For the nominalized form this will be a lemma, but without a full lemma article. It will have only the cross-reference to the main lemma. Examples are:

- Regular verb: पुल, པུལ། [pulup^{22}] *vt*. to push, shove. धकेल्नु, घचेट्नु
- Irregular verb: ल्हा, ལྷ། [lha^{2}] **pr**: ल्हेवी, **pc**: ल्हयी, **pd**: ल्हासुङ, **im**: ल्हो, **nom**: ल्हाप *vt*. 1) to look at. हेर्नु 2) to look after. हेरचाह गर्न.
- Nominalized form: ल्हाप, ལྷབ། [lhap^{2}] *cf*. nominalized form of ल्हा

5.2.3.1.3 Lemmatizing nouns

Sherpa nouns are relatively easier to be lemmatized compared to the situation in many other languages. The plural suffix (-tiwa), possessive suffix (-ki), and postposition marker (-la) are quite regular. Some exceptions are in the combination with a possessive suffix. For example, the noun possessive marker is –ki in the regular form:

- Regular form: दोकर + -की [dokar^{11} + -ki] ‘basket + of’

But there are some irregular forms created by morphological changes.

- Irregular form: दोङ्बु + -की [doŋbu^{11} + -ki] ‘tree + of’ → : दोङ्बी [doŋbi^{11}]

In this case the irregular form will be shown in the main lemma as **poss**: दोङ्बी. *poss* is an indicator of this possessive irregular form.

- Example of main lemma: **दोङ्बु**, རྡོང་བུ། [doŋbu[11]] **poss**: **दोङ्बी** *n*. a tree. **रुख**

The irregular form also will be a main lemma, but it is only with a cross-reference showing what the single form of the noun is.

- Example of irregular form: **दोङ्बी**, རྡོང་བི། [doŋbi[11]] *cf*. Poss. of **दोङ्बु**

The system of the lemmatizing nouns will be explained in more detail in the user's guide.

5.2.3.2 Different types of lemmata

Gouws and Prinsloo (2005:86) refer to the macrostructural selection of a dictionary as a collection of lexical items, and have classified these lexical items into three kinds: *words*, *items smaller than words*, and *items that consist of more than one word*. The types of lemmata will be decided by this classification. Different types of lemmata, which will be included in the Sherpa Dictionary are indicated below.

- One word items, and items consisting of more than one word
  1) Simplex words: Default form of a lemma
  2) Complex words such as a compound noun or verb. In the next section the kinds of compound verbs and nouns will be explained and it will be indicated whether they should be included as lemmata or sublemmata.
  3) Idioms
- Items smaller than words: Each case marker and particle will be a lemma.
  1) Case marker of ergative and possessive: -ki
  2) Case marker of dative, locative, instrumental: -la
  3) Case marker of ablative: nesur, etc.
  4) Particle: raŋ (emphatic), ti (emphatic), etc.

  Examples of case markers and particles:

  -की[1], ཀྱི། [-ki][69] *suf*. an ergative marker. **-ले**.

[69] The suffix itself does not carry tone.

-की[2], ཀི་ [-ki] *suf*. a possessive marker. -को.

नेसुर, ནས་སུར་ [nesur[11]] *pp*. from (one place to the other). देखि, बाट.

राङ, རང་ [raŋ[1]] *part*. an emphatic particle. नै.

5.2.3.2.1 Compound verbs as sublemmata

In Sherpa there will be two kinds of compound verbs, i.e. verb-verb and noun-verb.

- Verb-verb compound verbs: the first verb (Primary verb) does not carry any grammatical inflections, and is connected by the Connecter to the second verb (Light verb /Vector), which carries the grammatical variations. The meaning is carried by the primary verb, and the light verb provides only fine shades of meaning, which signify a sequence of actions. In this case, this compound verb is under the primary verb as a sublemma.

  Example: k^hun ɖop
  k^hu-n
  carry(PV)-Con go(LV): 'to carry'

- Noun-verb compound verbs: the meaning is determined by the noun, and the grammatical inflections are carried by the verb. In this case this compound verb is presented under the noun as a sublemma.

  Example: mo gyəkup
  oracle throw(a stone at): to tell somebody's fortune

5.2.3.2.2 Compound nouns as sublemmata

In Sherpa, there are two kinds of compound nouns: noun-adjective and noun-noun.

- Noun-adjective compound nouns: The noun carries a meaning as a head, and the adjective is a modifier that limits the meaning of the head. In this case, the compound noun is a main lemma without a space between the two words.
  Example: mikərwu (mik + kərwu, eye + white): 'a foreigner'

- Noun-noun compound nouns: The relationship of the two nouns is not attributive, but the meaning is a combination of the meanings of these nouns. The two nouns became a single word without a space. In this case, the compound noun is a main lemma.
  Example: mikcyur[70] (mik + c^{h}yu, eye + water): 'tear'

5.2.3.2.3 Idioms

Whereas the meaning of the compounds could be guessed by the individual words, the meaning of an idiom cannot be predicted, as the meaning is not related to the meanings of the individual words. As the idiom has a single meaning, it qualifies as a lemma, but it is not easy to put them in alphabetical order. The best way is to find out the key-word of the idiom and alphabetize them according to the order of that key-word as a sublemma (Gouws and Prinsloo 2005:88). In the Sherpa Dictionary, the key-word will be the key-noun of the idiom, i.e. lam (road), and k^{h}okpa (stomach) in the examples below will be the key-words. So, these idioms will be located under the lemmata of those key-words' as sublemmata.

- lam s^{h}ukup: 'to die'
  road to enter
  लाम शुकुप, ལམ་ཤུག་གུག [lam s^{h}ukup$^{1\ 11}$] *vi*. to die. मर्नु. ♦ती मी ती ला अलायी नावु तप्की लाम शुसुङ। He died after sickness of many years.

- k^{h}okpa lhap: 'to test (one's mind)
  stomach to see
  खोक्पा ल्हाप, ཁོག་པ་ལྷབ [k^{h}okpa^{11} lhap^{2}] *vt*. to fathom one's mind. मन चोर्नु. ♦ती मी तीकी खोक्पा ल्हापला ङला टङ्गा बडी बिन्सुङ। He gave a lot of money to test me.

5.2.3.2.4 Main lemmata and sublemmata: vertical vs. horizontal ordering

The sublemmata under the main lemma could be listed in a vertical order, or in a horizontal way to save the space. In the Sherpa Dictionary the sublemmata will be located under the main lemma 1) in a vertical order, 2) preceded by the indicator of a sublemma, which is an asterisk (*), 3) with the indentation of two spaces (one for the asterisk, and one empty space), 4) a tilde (~) is

[70] In Sherpa compound nouns, the onset of the second syllable has a tendency to lose the aspiration, if it is an aspirated consonant.

used when it is not in the main lemma[71], but only in pronunciation in order to save space as a textual condensation. One example of this is given below:

पेर्मी, བེར་མི། [permi[11]] *n*. a wife. बूढी, स्वास्नी

* पेर्मी छ्युङा, བེར་མི་ཆུང་བ། [~ c^hyuŋa[22]] *n*. a second wife (lit. a little wife). कान्छी स्वास्नी.

5.2.3.3 The ordering of lemmata

Traditionally there are two ordering systems in dictionaries, i.e. an alphabetical and thematic ordering. Gouws and Prinsloo (2005:96-97) reminded us that we should not underestimate the educational value of thematic ordering. However, in the Sherpa Dictionary, we do not use thematic ordering. In the following subsection, I will introduce the Sherpa alphabetical order, and the access alphabet, which was introduced by Nielsen (1995).

5.2.3.3.1 Alphabetical orders of the Sherpa language

The Sherpa orthography system generally follows the Devanagari alphabet system. However, the Sherpa language belongs to the Tibeto-Burman language group, some pronunciations do not fit the Devanagari script, which is generally used for languages in the Indo-Aryan language group. Therefore, some letters of the alphabet, which were adapted from the Devanagari, are unknown to most of the Sherpa people. So, this Sherpa alphabet system should be taught in local Sherpa schools and published for the people continuously in all kinds of teaching materials. The following tables are the vowel and consonant charts, and the newly modified letters of the alphabet are highlighted by putting them into boxes.

- The alphabetical order of the Sherpa vowels is: ə, a, i, u, e, o.

Sherpa phoneme	Initial symbol	Non-initial symbol
ə	अ	
a	आ	◌ा
i	इ	◌ि,◌ी
u	उ	◌ु
e	ए	◌े
o	ओ	◌ो

Table 14. Sherpa vowels

[71] As this Dictionary is for an endangered language, if we need a high degree of textual condensation, it is hard for the first learners. So, we don't use the tilde in the lemma, but only in the pronunciation.

- Alphabetical order of the Sherpa consonants is from top to bottom and left to right on each line as below, and the alphabet letters in the boxes are modified ones.

क	ka	ख	k^ha	ग	ga	ङ	nga				
च	ca	छ	c^ha	ज	ja	ञ्य	ɲa				
च्य	cya	छ्य	c^hya	ज्य	jya						
ट	ʈa	ठ	$ʈ^ha$	ड	ɖa						
त	ta	थ	t^ha	द	da	न	na				
प	pa	फ	p^ha	ब	ba	म	ma				
य	ya	र	ra	ऱ्ह	rha	ल	la	ल्ह	lha	व	wa
श	s^ha	स	sa	ह	ha						

Table 15. Sherpa consonants

5.2.3.3.2 The access alphabets in the Sherpa Dictionary

An access alphabet is not identical with the normal alphabet, but may contain additional codes or symbols, such as numbers, tone markers, and hyphens (Nielsen 1995). These access alphabets will be shown in the guidelines of the front matter texts to enable the first users to use the dictionary in a proper way. Gouws and Prinsloo (2005:98-100) introduced a typical starting point in the formulation of rules for an access alphabet as given below. The Sherpa Dictionary will follow these rules, if they are applicable.

- An unmarked form should always precede a marked form.
- An unhyphenated form precedes a hyphenated form.
- A post-hyphenated form precedes a pre-hyphenated form.

5.2.3.3.3 The homonyms in the Sherpa Dictionary

Homonyms involve two or more lexemes (lemmata) with identical phonological and orthographical forms but with no apparent meaning relationship. Traditionally the oldest lexical

item will precede the new lexical items in diachronic dictionaries (Gouws and Prinsloo 2005:100). In synchronic dictionaries, such as the Sherpa Dictionary, the historical perspective is not included, so here the frequency is applied rather than the history. Furthermore, in the Sherpa Dictionary, phonological and grammatical points are applied as indicated below:

- In regard to homonym pair, the word with high frequency precedes a word with less frequency.
- In regard to homonym pair, phonologically, a word with a low tone precedes a word with a high tone.
- In regard to homonym pair, grammatically, a pronoun precedes a noun, and a noun precedes a verb.

ङ ^{1}, ང་ ❸ [ŋa^{1}] *pron*. I. म.

ङ ^{2}, ལྔ ❸ [ŋa^{2}] *n*. five. पाँच.

ङ ^{3}, ཀྲ ❶[ŋa^{2}] *n*. a drum. ढोल, ढ्याङ्रो.

5.2.4 Microstructure of a Sherpa Dictionary

As mentioned in 5.2.3, the lemma and the whole set of data items, which are addressed to the lemma, form the dictionary article. The structural ordering of the data within the article is called the microstructure. Gouws (2014:160-161) stated that articles contain two types of text segments, i.e. items and indicators. Items are part of the microstructure; whereas indicators do not belong to the microstructure, but to the article structure. Items are data-carrying entries, e.g. pronunciation, morphology, part of speech, paraphrase of meaning, translation equivalents, illustrative examples, etc. Indicators, which are also known as structural indicators, are not data-carrying entries, but identify certain items or article slots.

There are two kinds of structural indicators, i.e. typographical and non-typographical structural indicators. Typographical structural indicators are the different typefaces, e.g. bold, italics, roman, and the use of capitals, small caps, etc. In the Sherpa Dictionary three types of typographical structural indicators will be used, i.e. bold, italics, and capital letters.

(1) Bold letters for verb paradigm markers

ल्हा, ལྷ། [lha^{2}] **pr**: ल्हेवी, **pc**: ल्हयी, **pd**: ल्हासुङ, **im**: ल्हो, **nom**: ल्हाप *vt*. 1) to look at. हेर्नु 2) to take care of. हेरचाह गर्न.

(2) Italic letters for the part of speech, cross-reference, and subject field (cf. 5.2.4.2.4) markers

छ्याङ्गा, ཚང་ག། [c^{h}yaŋga^{11}] (*rel*.) *n*. a death ceremony, which will be done 3 weeks after the death. मृत्यू संस्कार- जुन मृत्यू भएको गर्ने धर्मिक संस्कार. *syn*. ग्येवा

(3) Capital letters for the dialect variations (WD: western dialect, ND: northern dialect)

नम्ज्योक, ནམ་རྗོག [namjyok^{21}] *n*. an ear. कान. ND: अम्ज्योक.

Non-typographical structural indicators in the Sherpa Dictionary will occur in five places: sublemma, idiom, illustration, pronunciation, and synonym markers.

(1) Asterisk (*) for the sublemma indicator: Example is the same as in 5.2.3.2.4.

(2) Silcrow (§) for the idiom indicator

खोक्पा, ཁོག་པ་ [k^{h}okpa11] n. a stomach. पेट.

§ खोक्पा ल्हाप, ཁོག་པ་ལྷབ། [k^{h}okpa11 lhap2] vt. to fathom someone's mind. मन चोर्नु.

(3) Diamond (♦) for the illustration indicator

पड्बु, དཔང་པོ། [paŋbu^{22}] *n*. a testimony, witness. गवाही, साछी. ♦ती मी तीकी कुन कितुप बेला खोकी थुङ्गुप तप्की खो चोङ्खाङला पड्बु तेरुपला खोला कताङसुङ। He was summoned to the prison to give testimony, since he saw the person, when this person was robbed.

(4) Brackets ([ ]) for the pronunciation

खाङ्बा, ཁང་པ། [k^{h}aŋba^{22}] *n*. a house, building. घर, भवन

(5) Comma (,) for synonyms between the translation equivalents that are target language synonyms. In the example below, the two translation equivalents, i.e. house and building are synonyms.

खाङ्बा, ཁང་པ། [k^{h}aŋba^{22}] *n*. a house, building. घर, भवन

Gouws and Prinsloo (2005:119) stated that the article structure can be divided into two major article components, i.e. the comment on form and the comment on semantics. The comment on form is the part of the search area accommodating those data types that reflect on the form of the lemma sign, i.e. the morphological, phonetic, and orthographic form. The comment on semantics

contains the search zones accommodating those data types that reflect on the semantic and pragmatic features of the lexical item represented by the lemma (Gouws and Prinsloo 2005:125).

5.2.4.1 The comment on form in the Sherpa Dictionary

5.2.4.1.1 Variants

Among the three different dialects of the Sherpa language, there are two kinds of variants, i.e. those based on lexical differences and those based on phonological differences. The lexical differences among two or three dialects mean that the variants are lexically completely different. In this case, the variants will be shown by means of cross-references given under the lemma of the Southern dialect (cf. 5.1.1.4). The variant itself will also be included as a main lemma, which has a limited microstructural item, but it will be the guiding element of a cross-reference article, cross-referencing the user to the main lemma representing the Southern dialect, as illustrated below:

पेमा, འབྲས་ [pema[11]] *n*. sand, sandy soil. बालुवा, बलौटे माटो. WD: पेप्शोक
पेप्शोक *See* पेमा

If the variants are on the level of phonological differences, these differences will be indicated with regard to the specific lemma sign by means of the variant indicator of ND (Northern Dialect) or WD (Western Dialect). In this case the variant itself will not be included as a main lemma in its own alphabetical position. See the example below:

पुम, བུམ, ND: फुम [pum[1]] *n*. a daughter. छोरी.

5.2.4.1.2 Pronunciation and tone markers

In the Sherpa Dictionary, pronunciation will be indicated by means of brackets as the structural indicator. The phonetic representation will be by means of the International Phonetic Alphabet (the IPA). Since the Sherpa language has two-tone system, low and high, the tone of each syllable will be indicated with an item complementing the item giving the pronunciation as shown below:

low tone: [1] (superscript of 1), Example: [permi[11]]
high tone: [2] (superscript of 2), Example: [kʰaŋba[22]]

The Sherpa tone will be marked only in the dictionary and the primer, whenever the context does not accompany the word, because most Sherpa people could understand even minimal pairs according to the context in the narrative Sherpa text (Lee 2004).

5.2.4.1.3 Morphological data

As stated in 5.2.3.1.3, a few Sherpa nouns have irregular morphological changes, e.g., when the nouns have a possessive suffix. For those which have irregular changes, the comment on form will include entries indicating morphological data with a possessive indicator of **poss** as given below:

खाङ्बा, ཁང་བ། [kʰaŋba22] **poss**: खाङ्बी n. a house, building. घर, भवन

दोङ्बु, རྡོང་བུ། [doŋbu11] **poss**: दोङ्बी n. a tree. रुख

5.2.4.1.4 Part of speech

Within the comment on form the part of speech will be presented. It will be abbreviated in *italics*. The abbreviations used are listed below:

adj.: adjective
adv.: adverb
conj.: conjunction
dem. pron.: demonstrative pronoun
int. pron.: interrogative pronoun
n.: noun
part.: particle
post.: postposition
pron.: pronoun
suf.: suffix
vi.: verb intransitive
vt.: verb transitive

5.2.4.2 The comment on semantics in the Sherpa Dictionary

As mentioned in 5.2.4, the comment on semantics is the part of the search area accommodating those types of data that reflect on the semantic and pragmatic features of the lexical item represented by the lemma (Gouws and Prinsloo 2005:125). This component is mainly about the type of data giving a paraphrase of the meaning, and where the dictionary users want to consult with the dictionary, so the nature of the comment on semantics will be decided by the dictionary

type and dictionary user. In the next subsections, we will discuss the type and structure of the lexicographic definition, semantic labels, and cross-reference.

5.2.4.2.1 Lexicographic definitions

Both Zgusta (1971:252-253) and Landau (1984:153-154) start their definition of definition by making a distinction between logical definition and lexical definition. Philosophers such as Socrates wanted to define the concepts of the world, and the way people interpret these concepts. Zgusta and Landau's logical definition is to define an object by a definite contrast against everything else that is definable, positively and unequivocally; whereas the lexical definition focuses more on the semantic features of the lexical unit. The traditional rules of the lexical definition, based on Aristotle's analysis, define the word by identifying *genus* and *differentia*. First, find the genus, the species or family, and secondly, distinguish the word from all other things within that species or family. At this point Landau is correct that the lexicographer's job is not a theoretical exercise to increase the sum of human knowledge, but practical work to put together text that people can understand. Whereas the philosophers are concerned with the internal coherence of their system of definition; lexicographers are concerned with explaining something so their readers will understand. I agree with Landau, who defined the dictionary as an art and craft, so that the lexicography is a science of doing something useful for the people.

Zgusta mentioned that there is no one-to-one equivalence, and the whole area of application is divided differently between the two languages (1971:296). Since a definition started from the genus and the differentia, if the genus is different in the two languages of their cultures, the differentia could not stand on a firm foundation. In the next subsection, we will discuss the problem of anisomorphism, equivalence, and lexicographic definition, because this Sherpa Dictionary will be a typological hybrid offering both lexicographic definitions and translation equivalents.

5.2.4.2.1.1 Anisomorphism

Zgusta (1971:294) explained the purpose of the bilingual dictionary as the coordination of lexical units of one language with the corresponding lexical units of another language, and he added that a fundamental problem during this coordination is caused by the anisomorphism of languages. Anisomorphism is derived from a biological term, isomorphism, which means, similarity in

organisms of different ancestry resulting from evolutionary convergence[72]. In terms of equivalence relations, we are equating this anisomorphism with zero equivalence and the lexicographer treats this by means of a surrogate equivalent (cf. 5.2.4.2.1.2). Zgusta argued that two areas in lexicography where anisomorphism can occur.

(1) Culture-bound words

If some things exist only in the area of the source language, and not at all in the target language, there will be no real equivalent lexical units in the target language. The Sherpa culture is very similar to the Tibetan culture, so both Sherpa and Tibetan share cultural terms, especially religious expressions. However, it is quite hard to give an equivalent term in Nepali and even harder in English for these terms. For example,

> मिलु, མི་ལུས། [milu[11]] *n.* a shape or form of a life, which will be decided by the person's virtue in his former life. जीवनको आकृति वा रुप, जुन व्यक्तिको पूर्व जीवनको सद्गुणद्वारा निर्णय गरिनेछ. ♦ती मी तीकी दी मिजीला लाका मेलोवा क्यावु तप्की शिसिमा लाङगी मिलु लानी दुक्पा खुर गोकिवी। As this person has lived with bad works in this life, if he dies, he has to carry difficulties in the shape of an ox.

The Sherpa Dictionary will paraphrase the word that are anisomorphisms in the target language, and give cotext entries as an illustration to show the meaning clearly.

(2) Grammar-bound words

If one aspect of grammar in the source language is not able to be applied in the grammar of the target language, we have to explain the grammatical function with an illustration. For example, in the Sherpa pronominal system, there is an exclusive function in the first plural pronoun, which is not the case in English or Nepali.

> ङियराङ, ངེ་རང་[ŋiraŋ[11]] *pron.* a pronoun of first person plural with exclusive. पहिलो पुरुषको बहुबचनको सर्वनाम (हामी), तेस्रो पुरुष समावेश नभएको. ♦खो ताङ मुला ङियराङ जो ल्यामु क्यासिनङ ती मीतिवा छासे ङ्योशोक ङ्योशोक सिनोक। Even though he and we (Excl.) did good things, they complained (lit. do murmuring murmuring) a lot.

[72] From the Medical Dictionary of Merriam-Webster. http://www.merriam-webster.com/dictionary/isomorphism [September 27, 2016]

5.2.4.2.1.2 Equivalence

Zgusta defined equivalence as a lexical unit of the target language which has the same lexical meaning as the respective lexical unit of the source language, and added that the lexicographer's most important duty is to find in the target language such lexical units as are equivalent to the lexical units of the source language, and to coordinate the two sets (Zgusta 1971:312). He and Gouws & Prinsloo introduced the following types of equivalences (Zgusta 1971:312-325, Gouws and Prinsloo 2005:154-163).

(1) Full or absolute equivalence

Full equivalence prevails when the lexical items of both the source language and the target language have exactly the same meaning, function on the same stylistic level and represent the same register. Even though the generic names in Sherpa and Nepali are mostly identical, this Sherpa Dictionary is for an endangered language, so example sentences will be provided. Such illustrations will be beneficial to the readers, who will use this dictionary for text production.

> च्यच्युम, ཅྱ་ཅུམ། [cyəcyum[11]] *n*. a generic term for a bird. चरा. ♦मेराङ दोङ्बु थुक्पु वोतुप तप्की मेराङ दोङ्बी गोला च्यच्युम शासा देकिनोक। Since the pine tree grew densely, a bird made its nest on top of the tree.

(2) Partial equivalence

Gouws and Prinsloo (2005:155-158) stated that partial equivalence prevails when the source and target language items do not display a one-to-one relationship. A frequent type of partial equivalence sees the establishment of a one to more than one relation between the source language and the target language. Gouws and Prinsloo defined this one-to-more-than-one relation as divergence. They divided this divergence into two subtypes, i.e. lexical divergence and semantic divergence.

a) Lexical divergence

Lexical divergence comes, when the monosemous lexical item of the source language has more than one translation equivalent in the target language. In most cases, these equivalents are partial synonyms in the target language. The comma (,) will be used as a structural indicator to show that the translation equivalents connected by the comma are synonyms. For example, in the Sherpa example below, the comma between *quickly* and *rapidly* indicates that these two translation equivalents are synonyms.

ङमु, ངྨུ། [ŋəmu[11]] *adv*. quickly, rapidly. छिटो, चाँडो. ♦हारिङ जाँच वोतुप तप्की खो ङमु लोप्टाला गाल्सुङ। Because of the test today, he went to school quickly.

b) Semantic divergence

Semantic divergence comes, when the lemma sign represents a polysemous lexical item. In this case, each polysemous sense and the translation equivalent will be entered as a subcomment on semantics. For the communicative equivalence, each subcomment on semantics will have illustrations to show the contexts (Gouws and Prinsloo 2005:161-163). Regarding types of microstructure, the Sherpa Dictionary will use the integrated microstructure, so that each example sentence will come immediately after the translation equivalent respectively. In the non-integrated microstructure, all the example sentences belonging to each translation equivalent will be separated into a text block. The following is an example of integrated microstructure in Sherpa.

ओङ, དབང་། [woŋ[1]] *n*. 1) power. शक्ती. ♦खो ति नाज्युङ वोतुप बेला च्यालक च्येनदी साङ खुन डोप ओङ साङ नोक, तन्दा ति मुथुप्नोक। When he was young, he had the power to carry even a heavy load, but now he cannot do it. 2) authority. अधिकार. ♦खो ति ती युलकी मी छ्ये यिन्दुप तप्की ती टङ्गा लङ्गुप ओङ वे। As he is the chief of the village, he has the authority to receive the money.

(3) Zero equivalence

Gouws and Prinsloo (2005:158-160) stated that zero equivalence prevails where the target language has no item that can be coordinated as a translation equivalent with a lemma representing a source language item. Zero equivalence often leads to the inclusion of surrogate equivalents (cf. 5.2.4.2.1.1), i.e. a target language entry substituting a translation equivalent. The Sherpa language, which has a strong influence from Tibetan culture, has many lexical items for things or concepts, which do not exist in Nepali and English. In this case, the meaning should be paraphrased in the target language. One example from religious terms is given below:

छ्योर्तेन [c^hyorten[11]] (*rel*.) *n*. a dome-shaped tower of Tibetan Buddhist shrine, Chorten. गुमज आकारको भोटेहरुको देवमन्दिर. ♦हारिङ छेवा च्येङा यिन्दुप तप्की खोतिवा छ्योर्तेन कोरा ग्यकुप गाल्सुङ। Today is the full moon day (lit. 15th lunar day), so they went for going around the Chorten.

5.2.4.2.2 Context and cotext entries

Gouws and Prinsloo (2005:127) stated that in a dictionary compiled for text reception exclusively it is acceptable to limit the treatment presented in the comment on semantics to the mere presentation of a paraphrase of the meaning or a translation equivalent. However, for text production, the lemma sign and the translation equivalents in active communication should be used to help the user to use the word properly. For this purpose, the relevant context or cotext will be good to be included in the lemma and translation equivalents. The context is usually glosses to show the usage of the word in the pragmatic environment; whereas the cotext consists of illustrative examples to indicate the syntactic environment. The following are some Sherpa examples.

(1) Context: glosses

> **कोरा ग्यकुप**, སྐོར་ར་རྒྱག་གུབ། [kora gyakup[2211]] *vi*. 1) to go around (esp. a Chorten, a Tibetan Buddhist tower, for the purpose of earning religious merit). **परिक्रमा गर्नु. ♦हारिङ छेवा च्येङा यिन्दुप तप्की खोतिवा छ्योर्तेन कोरा ग्यकुप गाल्सुङ।** Today is the full moon day (lit. 15th lunar day), so they went for going around the Chorten. 2) to travel (from one place to another). **यात्रा गर्नु. ♦दाक्पी लुङ्बाला कङ्री नछोक वोतुप तप्की छ्यिग्यपकी मीतिवा दाक्पी लुङ्बा कोरा ग्यकुपला गिवी।** In our (Inc.) country since there are many high mountains, many foreign people come to our (Inc.) country to travel.

(2) Cotext: illustrative examples

> **कोन्दुप**, གོན་དུབ། [kondup22] *vt*. to put on. **(लुगा) लगाई दिनु. ♦नाम टङ्गा वोतुप तप्की ती अम्पुम तीकी खोरो आङला मज्या टेन्बु कोन बिन्सुङ।** Since the weather became cold, that lady put warm clothes on her baby.

5.2.4.2.3 Collocations

The term collocation refers to sequences of lexical items, i.e. habitually co-occurring lexical items or mutually selective lexical items (Cruse 1986), but there are different definitions according to the function of the combination. Here, two functions will be explained: restrictedness and transparency.

5.2.4.2.3.1 Collocations and restrictedness

A collocation can be defined by its degree of restrictedness, such as unrestricted, semi-restricted, and restricted collocation. In an unrestricted collocations, it is possible for the word to be used in combination with many other words, e.g. take as in: *-take a look / a holiday / a rest / a letter / time / notice / a walk/ a notion.* However, a restricted collocation is a fixed combination of words, which is not open to combination with other words, e.g. *dead drunk, pretty sure, stark naked, consider seriously, lean meat, accept defeat*. Cowie (1978) defined it similarly as restrictedness vs. openness. Openness here has the same function as unrestrictedness. Restrictedness is a measurement to contrast the collocations from an open phrase. In the Sherpa Dictionary, unrestricted collocations will not be treated as collocations. Sherpa examples of restrictedness are as follows:

- unrestrictedness*:* mi (man) / tep (book) toŋgup 'to send'
- restrictedness: nasam toŋgup : 'to think'
  thought to send

  c^hyowa toŋgup : to read scriptures in the ceremony
  scripture to send

5.2.4.2.3.2 Collocation and transparency

Cruse (1986) also defined collocations in terms of the degree of transparency. Let's say that there are two vocabularies A and B. And the meaning of A is *A*, and the same of B is *B*. If the meaning of A+B is *A+B*, they are an open phrase, because they are transparent to include both meanings. If the meaning of A+B is *A+BB* or *AA+B*, they are collocations, because basically they are transparent to possess half of the meaning. And if the meaning of A+B is C, which is not transparent at all, they are idioms. Sherpa examples are:

- Open phrase: k^haŋbi naŋla s^hukup : 'to go into the house'
  house-Gen inside to enter
- Collocation: ŋyasala s^hukup : 'to go to bed'
  bedroom-Loc to enter
- Idiom: lam s^hukup : 'to die'
  road to enter

5.2.4.2.3.3 The categories of Sherpa collocations

Collocation consists of a base and a collocator. The meaning of the base is independent, and the meaning of the whole collocation becomes clear only by means of the collocator. Let's give some examples from English: *to play the piano, to play tennis, to play cards*. The nouns here are the base, and the verbs, the collocator. The meanings of the nouns are not dependent, but will be made clear with the verbs as collocators.

The reason why we have to analyze this collocation is to know the place, where this collocation is located, whether in the base or in the collocator. So, we will discuss more about the categories of Sherpa collocations to decide the location of the collocation for each category. Sherpa collocations can be distinguished by the grammatical categories to which they belong. In each category, the location of collocation will be shown.

- NOUN[73] + VERB: the noun is the base and the verb is the collocator. In this category, the collocation will be located in the noun as a sublemma.
  - no agent marker: gaŋpa s^{h}arup : 'to wet the bed'
    bladder to break
  - agent marker: kaŋba/i[74] ɖop : 'to go on foot'
    foot-Ag to go
  - locative marker: lo-la jyokup : 'to keep in mind'
    mind-Loc to put
- NOUN + NOUN: The first noun is a base, and the second is a collocator. In this case, the collocation will be on the first noun as a sublemma.
  - ŋyimi kuŋ : 'noontime'
    day+Gen center
  - caŋbi gari : 'river bank'
    river+Gen edge
- NOUN + ADJECTIVE : the noun is the base, and the adjective is the collocator. In this case, the collocation will be located in the noun.

[73] The nouns either will haveor will not have a case marker.
[74] kaŋba will be changed by the Genitive marker (-ki) to become kaŋbi

- go yewu [75]: 'be perplexed'
  head be hang around
- ŋo nakpu : 'unhappy'
  face black

- ADVERB + VERB : The adverb is the collocator, and the verb is the base. In this case, the collocation will be located in the verb.
  - tʰəratʰura toŋgup : 'to make separately'
    scattered to send
  - dokolok ɖop : 'to be blunt'
    blunt to go

5.2.4.2.4 Lexicographic labels

Gouws and Prinsloo introduced lexicographic labels as pragmatic markers to relate an item in a dictionary to the world outside the dictionary. If we use these markers for a lemma sign, all senses of the word fall within the scope of the label. In the Sherpa Dictionary, two kinds of labels will be used, i.e. subject field labels and stylistic labels.

(1) Subject field labels

Zgusta called the subject field labels as 'classificatory labels' (Zgusta 1971:332). It would be ideal to collect the data of lemmata from an early stage by means of labelling the semantic domains, so that each lemma has its own semantic domain. Good examples of semantic domains are Moe's list (Cf. 4.3.3) or Murdock's Outline of Cultural Materials Subjects List (Murdock 1987). The labels will be three-letter abbreviations in italics, and their full lists will be shown in the guidelines of the dictionary. Examples from the lists and their applications in the Sherpa Dictionary are shown below:

ani(mal)
bio(logy)
foo(d)
geo(graphy)
kin(ship)
rel(igion)
too(ls)

[75] yewu is a nominalized verb as an adjective.

पालाङ, བ་སྤང་ [palaŋ[11]] (*ani.*) *n.* a cow. गाई

(2) Stylistic labels

Stylistic labels are used to mark deviations from the standard variety, neutral register, and style of the language, e.g., honorific, colloquial, coarse language, and slang. In this case, an illustration will be added as a comment on semantics. The labels also will be three-letter abbreviations in italics, and these labels will be shown in two places, i.e. 1) after the pronunciation to show the style of the lemma, 2) in the translation of the example sentence to mark the style of the lemma. Examples of the lists and the applications in the Sherpa Dictionary are given below:

coa(rse language)
col(loquial)
hon(orific)
inf(ormal)
sla(ng)

फेपुप, ཕེབས་ཤུག [p^{h}epup[11]] (*hon.*) *vi.* to come. आउनु. ♦दाङ ङे पालु ताङ आमा ति अस्पताल डोपला युल नेसुर यम्बुरला फेप्सुङ। Yesterday my parents came (*hon.*) to Kathmandu from the village for going to the hospital.

5.2.4.2.5 Cross-reference: Mediostructure

Gouws and Prinsloo (2005:177) stated that the cross-reference can be used to establish relations between different components of a dictionary. Wiegand (1996a:11) added that it interconnects the knowledge elements represented in different sectors of the dictionary on several levels of lexicographic description to form a network. Gouws and Prinsloo called this cross-reference mediostructure. The cross-reference entry consists of a cross-reference marker and a cross-reference address. The kinds of cross-reference markers and their applications in the dictionary will be as follows:

(1) Synonym

छ्याङ्गा, ཆང་ཀ། [c^{h}yaŋga[11]] (*rel.*) *n.* a death ceremony, which will be done 3 weeks after the death. मृत्यू संस्कार- जुन मृत्यू भएको गर्ने धर्मिक संस्कार. *syn.* ग्येवा

(2) Antonym

येङ्गा, ཡང་ག [yeŋga[11]] *adj.* light. हलका. *ant.* च्येन्दी

(3) Variation: If a lemma has a variant, it will be expressed in the cross-reference. The variation itself does not have a full dictionary article, but the reader will be guided to look up the main lemma by the variation marker *See*.

मेलोक्पा, མེ་ལོག་པ། [melokpa[11]] *adj*. 1) bad. नराम्रो. 2) wicked. खराब. *ant*. ल्येमु *See* मेलोवा

(4) Words with a close relationship, e.g. key-lock, bow-arrow, and wide-high, will be shown for the further information of the readers by the marker, *cf*.

गोल्ज्याक, སྒོ་ལྕགས། [goljyak[11]] *n*. lock. ताल्चा. *cf*. लिमी

5.2.4.3 The addressing structure

Gouws and Prinsloo (2005:134-135) stated that all lexicographers should pay attention to the scope of each entry, because all entries are functional either as presenting lexicographic data or structural indicators identifying data categories. The lemmata are included in the macrostructure of a dictionary as the guiding elements of the articles, so they will be the primary or first-level treatment units. Each microstructural item is part of the treatment of either the lemma or of another microstructural item, which means that microstructural items are directed or addressed at specific targets. With this as in the background, Gouws and Prinsloo defined the addressing structure as follows: "The addressing structure of a dictionary is the system according to which these procedures of one item being directed at another is employed." The lemma is the most typical address, but other items also function as addresses. This results in two types of addressing, i.e. lemmatic and non-lemmatic addressing.

5.2.4.3.1 Lemmatic vs. non-lemmatic addressing

According to Gouws and Prinsloo (2005:135) the lemmatic addressing is a procedure where the main lemma is the address of an entry. In a dictionary, since all lemmata are in alphabetical order and will be arranged vertically, the lemma will be the guiding element of an article. Non-lemmatic addressing is a procedure where the lexicographic treatment is directed to an item not functioning as a lemma. Here, we have to be reminded that whereas lemmatic addressing is directed at macrostructural items, non-lemmatic addressing is addressed at microstructural items. The importance of this address is that the address is the topic of the specific treatment procedure. When we use non-lemmatic addressing, it implies a system of topic switching within the

dictionary article. Gouws and Prinsloo (2005:135) stated that especially the bilingual dictionaries have traditionally been dominated by a lemmatic addressing bias. Because of this bias, the translation equivalents were working just as part of the treatment of the lemma, and it was hard to add data addressed at the translation equivalents in order to help the user to choose a correct translation equivalent. At this point, we have to say that the translation equivalents should also function as secondary treatment units in a dictionary article.

5.2.4.3.2 The addressing structure in the Sherpa Dictionary

The envisaged Sherpa Dictionary will display both a lemmatic and a non-lemmatic addressing. The level of equivalence will be considered, but, because this dictionary is for an endangered language, cotext and context will be provided (cf. 5.2.4.2.1.2). An example in the Sherpa Dictionary is:

> कोरा ग्यकुप, རྐོར་ར་རྒྱག་གུབ། [kora gyakup[2211]] *vi*. 1) to go around (esp. Chorten, Tibetan Buddhist tower for the purpose of earning religious merit) ♦हारिङ छेवा च्येङा यिन्दुप तप्की खोतिवा छ्योर्तेन कोरा ग्यकुप गाल्सुङ। Today is the full moon day (lit. 15th lunar day), so they went for going around the Chorten. 2) to travel (from one place to another) ♦दाक्पी लुङ्बाला कङ्री नछोक वोतुप तप्की छ्यिग्यपकी मीतिवा दाक्पी लुङ्बा कोरा ग्यकुपला गिवी। In our (Inc.) country since there are many high mountains, many foreign people come to our (Inc.) country to travel.

5.3 Chapter summary

The purpose of this thesis is to provide a theoretical model for dictionaries for endangered languages. To accomplish this purpose, first, in this chapter, I revisited the theories of meta-lexicography and confirmed a conceptualization of a dictionary compilation for the Sherpa language in 5.1. And then, as a lexicographer compiling a dictionary for an endangered language, I tried to apply all the theories in planning the model of the envisaged Sherpa Dictionary. In 5.2, the outer texts (both front matter texts and back matter texts) were explained, and then the macro- and microstructure of the central list were discussed. I always tried to keep two things in mind, the first was to connect the theories to the user's perspective, so that the user could retrieve the information relevant to fulfilling the genuine purpose of the Sherpa Dictionary. Second, I

wanted my applications to provide guiding principles for people wanting to create a dictionary for other endangered languages. That is also an important purpose of this thesis.

Chapter 6. Conclusion

6.1 Conclusion

This thesis is about endangered languages, how to revitalize them, and what role the dictionary plays in documenting these dying languages. The Ethnologue (Lewis et al. 2015) reports that there are 7,102 living languages in the world, but the saddest information is that "about 90% of the languages may be replaced by dominant languages by the end of the 21st century (UNESCO 2003:2)". Since this world is changing so rapidly, and it is so vital for all societies to remain in contact with other societies, we cannot stop the contact of their languages that will result in the minor languages becoming endangered as Thomason (2015:11) explained. At this point, the most important thing to do when a language is down to a few speakers is to document the knowledge of those speakers as thoroughly as possible (Hinton 2001:413). There will be many ways to help in language documentation, but the compilation of a dictionary is a crucial form of language documentation to revitalize a language. A dictionary is a good tool for outsiders to learn the language, and for insiders to document their language. This dictionary is particularly important because it is not only for the present generation but it will be for generation after generation.

The purpose of this thesis is to show the envisaged Sherpa Dictionary as a model for how to plan and compile a dictionary for endangered languages. For this purpose, in Chapter One of this thesis, I started by giving attention to the situation of endangered languages in the world, and focused on the importance of compiling dictionaries to revitalize the languages. In Chapter Two I explained the situation of the Sherpa language as a sample case, which has a few problems in regard to documentation. These problems would be very similar for other endangered languages, i.e. dialects, variations, and orthography. All these aspects present challenges for standardization. In Chapter Three I discussed some aspects of lexicography theories, i.e. typology, the theoretical approach to standard-preserving dictionaries, the general theory of Wiegand, and the function theory. For each topic, I tried to apply the theory to the Sherpa Dictionary. In Chapter Four the data collection was treated. However, before getting involved in data collection, the language study should be done; otherwise all data will be a mixture of different dialects and variants, which will be a big headache for the lexicographers later on in the real process of compiling a dictionary. After the dialect map is drawn, a long process of data collection will be started. I

discussed mainly two kinds of data collection, i.e. data collection by the corpus and data collection by semantic domains. The corpus method is the best one to use in data collection, because it gives the present-day, natural usage, but it takes time and money. The gaps in the corpus can be filled by the semantic domain method. These two methods together can complement each other. In Chapter Five I discussed the structures of the Sherpa Dictionary as a model for the endangered languages. First, I reorganized the typology and function of the Sherpa Dictionary. Then, I described the structure of the Sherpa Dictionary, i.e. the outer texts (front matter and back matter), and the central list using macro- and microstructure.

As a field lexicographer, I prepared this thesis in order to help other lexicographers or lay people, who are interested in compiling a dictionary for endangered languages, to see an example of how to plan such a dictionary. As a bit of personal history, the Sherpa Dictionary is my second dictionary, having already compiled the Nepali-Korean Dictionary (Lee 1999). When I started the Nepali Dictionary, I did not have any education in lexicography other than general linguistic knowledge, and I remember that I spent some quite difficult times trying to figure out the system of dictionary compiling, which is considered a dictionary conceptualization as I now understand it. I wouldn't have wanted to embark again on such a trial and error way in beginning to work on the Sherpa Dictionary. I, therefore, hope this thesis can serve a guideline for people who are interested in compiling dictionaries for endangered languages.

6.2 Further study needed

The main purpose of this thesis, in the final analysis, is that it will bring about the standardization of the Sherpa language. This thesis itself is also one part of the standardization process. For revitalizing endangered languages, standardization is one of the most important steps. Without it, the data collection and the analysis of the data would be in vain. The standardization can be done by linguistic methods, but more important is the role of the language society itself and their decisions in this regard. Generally, for the society represented by the endangered language, the topic of the standardization of their language is not a high and urgent priority, if it does not provide enough food for their hunger. However, if the language society does not agree with any language policy that has been decided by outsiders, any work done is in vain. So waiting for the time until the language-speakers understand the problem and accept the decision as theirs is most crucial. Sometimes this is a painful time for the lexicographers, and it

also means losing project money. However, if the dictionary compiler omits this course of action because of the difficult process of the standardization, later on he/she has to pay a much higher price. Furthermore, this dictionary is the property of the language society itself. Personally, I have been waiting for the standardization of the Sherpa language since the first meeting for the standardization of the Sherpa language on April 1st, 2002. I think one conclusion of this thesis could be that the dictionary is for the people, the readers. All processes, types, functions, and structures should be for the readers and be valuable to them from their perspective. Together with waiting for the Sherpa people's understanding of the standardization of their language, the specific lexicographic study of this topic should be continued by writing a paper on the study of the theory and observation of other cases of endangered languages.

Reference list

Al-Kasimi, A.M. 1977. *Linguistics and Bilingual Dictionaries*. Leiden: E. J. Brill.

Beekman, J. 1975. Eliciting Vocabulary, Meaning, and Collocations. In: Healey, A (Ed.). 1975. *Language Learner's Field Guide*. Ukarumpa, Papua New Guinea: Summer Institute of Linguistics.: 361-388.

Bergenholtz, H. and S. Tarp. 2003. Two opposing theories: On H.E. Wiegand's Recent Discovery of Lexicographic Functions. In: *Hermes, Journal of Linguistics*. 31: 171-196

Bergenholtz, H., S. Tarp and H.E. Wiegand. 1999. Datendistributionsstrukturen, Marko- und Mikrostrukturen in neueren Fachwörterbüchern. In: Hoffmann, L. et.al. (Eds.) *Fachsprachen. Languages for Special Purposes. An International Handbook of Special-Language and Terminology Research.* Berlin, De Gruyter: 1762-1832. Quoted in: Gouws and Prinsloo (2005):5, 58.

Bernard, H. R. 1996. Language Preservation and Publishing. In: Hornberger, N.H (Ed.). *Indigenous Literacies in the Americas: Language Planning from the Bottom up*. 139-156. Berlin: Mouton de Gruyter. Quoted in: UNESCO Ad Hoc Expert Group on Endangered Languages, Language Vitality and Endangerment (2003):2.

Blair, F. 1990. *Survey on a Shoestring: A manual for small-scale language surveys.* Dallas, TX: Summer Institute of Linguistics and the University of Texas at Arlington. Quoted in: Lee (2003):85

Casad, E. 1974. Dialect Intelligibility Testing. (Summer Institute of Linguistics Publications in Linguistics and Related Fields, 38) Norman, Oklahoma: The Summer Institute of Linguistics of the University of Oklahoma. Quoted in: Lee (2003):88.

Central Bureau of Statistics. 2011. *Statistical Pocket Book.* Central Bureau of Statistics.

Cowie, A.P. 1978. The place of illustrative material and collocations in the design of a learner's Dictionary. In: Stevens. P (Ed.). *In Honour of A.S. Hornby.* Oxford Uni. Press.

Cruse, D.A. 1986. *Lexical Semantics.* Cambridge University Press

Duda, W. 1986. Ein >>aktives<< russisch-deutsches Wörterbuch für deutsch-sprachige Benutzer?. In: Günther E (Ed.). *Beiträge zur Lexikographie slawischer Sprachen*. Berlin: Akademieverlag. 9-15.

Gallardo, A. 1980. Dictionaries and the Standardization Process. In: Zgusta, L (Ed.). 1980. *Theory and Method in Lexicography*. Columbia: Hornbeam Press:59-69.

Gordon, K.H. 1969. Sherpa Phonemic Summary. In: *Tibeto-Burman Phonemic Summaries,* Vol.VII. Kathmandu: Summer Institute of Linguistics and Tribhuvan University.

Gordon, K.H. and B. Schöttelndreyer. 1970. Sherpa segmental synopsis. In: Hale, A. and K. Pike (Eds.). *Tone systems of Tibeto-Burman languages of Nepal*. Urbana, IL: The University of Illinois. Part I. 345-357.

Gouws, R.H. 1996. Bilingual Dictionaries and Communicative Equivalence for a Multilingual Society. In: *Lexikos*. 6:14-31.

Gouws, R.H. 2001. Lexicographic training: Approaches and topics. In: Emejelu, J. du P (Ed.). *Elements de Lexicographie Gabonaise*. Jack Hillman Publishers:58-94.

Gouws, R.H. 2014. Article Structures: Moving from Printed to e-Dictionaries. In: *Lexikos*. 24:155-177.

Gouws, R.H. and D.J. Prinsloo. 2005. *Principles and practice of South African lexicography.* Stellenbosch, South Africa: Sun Media.

Gouws, R.H., W. Schweickard, and H.E. Wiegand. 2013. Lexicography through the ages: From the early beginnings to the electronic age. In: Gouws, R. H. et al. (Eds.) 2013. *Dictionaries. An International Encyclopedia of Lexicography*. Supplementary Volume: Recent Developments with focus on Electronic and Computational Lexicography. Berlin: De Gruyter:1-24.

Greninger, D.E. Another look at Storyline marking in Sherpa narrative. In: Hildebrandt, K. et al. (Eds.). 2011. *Himalayan Linguistics*. Vol. 10(1). 77-99. [Online]. Available: http://escholarship.org/uc/item/3qn376sn (accessed September 2015).

Grenoble, L.A. and L.J. Whaley. 2006. *Saving Languages: An Introduction to language revitalization.* Cambridge: University Press.

Hale, A. 1970. A Phonological Survey of Seven Bodic Languages of Nepal. In: *Occasional Papers of the Wolfenden Society on Tibeto-Burman Linguistics* Vol.III. 1-33. Publications of the Department of Linguistics. Urbana, IL: The University of Illinois.

Hale, K. On endangered languages and the safeguarding of diversity. In: Hale, Ken. et al. (Eds.). *Endangered Languages*. Linguistic Society of America. 1992: 1-3. [Online]. Available: http://www.jstor.org/stable/416368 (accessed August 2015).

Hartmann, R.R.K. 1989. Sociology of the Dictionary User: Hypothesis and Empirical Studies. In: Hausmann, F.J., et al. (Eds.). 1989-1991:102-111.

Hausmann, F.J. 1977. *Einführung in die Benutzung der neufranzösischen Wörterbücher*. Tübingen: Niemeyer. Quoted in : Tarp (2008):22.

Hausmann, F.J. 1989. Kleine Weltgeschichte der Metalexikographie. In: Wiegand, H.E (Ed.). *Wörterbücher in der Diskussion. Vorträge aus dem Heidelberger Lexikographischen Kolloquium*. Tübingen: Niemeyer. 75-109. Quoted in: Tarp (2008): 15, Gouws and Prinsloo (2005):45.

Hausmann, F.J. et al. (Eds.). 1989. Wörterbücher. Dictionaries. Dictionnaires. An International Encyclopedia of Lexicography. Berlin: De Gruyter.

Hausmann, F.J. and H.E. Wiegand. 1989. Component Parts and Structures of General Monolingual Dictionaries: A Survey. In: Hausmann, F. J., et al. (Eds.). 1989-1991: 328-360.

Hinton, L. 2001. Sleeping Languages: Can they be awakened? In: Hinton, L. and K. Hale (Eds.). *The Green Book of Language Revitalization in Progress*. Academic Press. 2001.

Kammerer, M. and H.E. Wiegand. 1998. Über die textuelle Rahmenstruktur von Printwörterbüchern: Präzisierungen und weiterführende Überlegungen. In: *Lexicographia* 14: 224-237. Quoted in: Gouws and Prinsloo (2005):5-6, 57, 59.

Kelly, B. 2003. A grammar and glossary of the Sherpa language. In: Genetti, C (Ed.). *Tibeto-Burman Languages of Nepal: Manange and Sherpa*. Canberra: Pacific Linguistics. 244-452.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London & New York: Longman.

Krauss, M. The world's languages in crisis. In: Hale, K. et al. (Eds.). 1992:4-10. *Endangered Languages*. Linguistic Society of America. [Online]. Available: www-01.sil.org/~simonsg/preprint/Wisconsin Symposium.pdf (accessed August 2016).

Kromann, H.P. et al. 1991. Principles of Bilingual Lexicography. In: Hausmann, F.J. et al. (Eds.). 1989-1991:2711-2728. Quoted in: Gouws (1996):18.

Ladefoged, P. Another view of Endangered Languages. In: Linguistic Society of America. *Language*. Vol.68, No. 4.(Dec., 1992), 809-811. [Online]. Available: http://links.jstor.org/sici?sici=0097-8507%28199212%2968%3A4%3C809%3AAVOEL%3E2.0.CO%3B2-5 (accessed August 2015).

Landau, S. 1984. *Dictionaries: The Art and Craft of Lexicography*. New York: The Scribner Press.

Lee, S.Y. 1999. *Nepali-Korean Dictionary*. Kathmandu: Hisi Printers.

Lee, S.Y. 2002. A Survey for script decision for the Sherpa-Nepali-English-Tibetan Dictionary. Paper presented at the 22nd Annual meeting of the Linguistic Society of Nepal. (Unpublished manuscript)

Lee, S.Y. 2003. A sociolinguistic survey of Sherpa. In: Kansakar, T.R. and M. Turin (Eds.). *Themes in Himalayan Languages and Linguistics*, 81-95. Heidelberg: South Asia Institute; Kathmandu: Tribhuvan University.

Lee, S.Y. 2004. Do tone markers help readability? Paper presented at the 24th Annual meeting of the Linguistic Society of Nepal. (Unpublished manuscript)

Leech, G. 1991. The state of the art in corpus linguistics. In: Aijmer and Altenberg 1991: 8-29. Quoted in: Kennedy (1998):62.

Lewis, M.P., G.F. Simons, and C.D. Fennig (Eds.). 2015. *Ethnologue: Languages of the World*, Eighteenth edition [Online]. Available: https://www.ethnologue.com/endangered-languages (accessed August 2015)

Lewis, M.P., G.F. Simons. 2010. Assessing Endangerment: Expanding Fishman's GIDS. [Online]. Available: http://www.lingv.ro/resources/scm_images/RRL-02-2010-Lewis.pdf (accessed August 2015)

Louw, J.P. and E.A. Nida. 1988. *Greek-English Lexicon of the New Testament Based on Semantic Domains*, Vol I. United Bible Societies.

McEnery, T. and A. Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: University Press.

McEnery, T. and M. Ostler. 2000. 'A new agenda for corpus linguistics: Working with all of the world's languages'. *Literary and Linguistic Computing* 15 (4): 403-30. Quoted in: McEnery (2012):12.

Mdee, S.J. 1999. Dictionaries and the Standardization of Spelling in Swahili. In: *Lexikos*. 9: 119-134.

Mikkelsen, H.K. 1992. What did Scerba Actually Mean by Active and Passive Dictionaries? In: Hyldgaard-Jensen, K. and A. Zettersten (Eds.). *Symposium on Lexicography V.* Proceedings of the Fifth International Symposium on Lexicography May 3-5. 1990, at the University of Copenhagen. Tübingen: Niemeyer: 25-40. Quoted in: Tarp 2008:20.

Ministry of Education in Nepal, Dept. of Education, Flash I Report 2068 (2011-2012). [Online]. Available: http://www.ncf.org.np/upload/files/1006_en_flash%20i%202068%20(2011-12)_1346396154.pdf

Moe, R. 2001. Lexicography and mass production. *Notes on Linguistics* 4.3: 150-156. Summer Institute of Linguistics

Moe, R. 2003. Compiling Dictionaries Using Semantic Domains. In: *Lexikos*. 13:215-223.

Murdock, G.P. and S.F. Clellan. et al. (Eds.). 1987. Outline of cultural materials Subjects Lists. Human Relations Area Files, Institute of Human Relations. Yale University. Available: www.ingramanthropology.com/uploads/6/8/1/1/6811328/ocm.pdf (accessed May 2016).

Nagono, Y. 1980. *Amdo Sherpa Dialect: A Material for Tibetan Dialectology.* Institute for the Study of Languages and Cultures of Asia and Africa.

Newell, L.E. 1995. *Handbook on Lexicography: For Philippine and Other Language.* Summer Institute of Linguistics. Linguistic Society of the Philippines, Manila.

Nielson, S. 1995. Alphabetic macrostructure. In: Bergenholtz, H. and S. Tarp (Eds.). 1995:190-195. Quoted in: Gouws and Pinsloo 2005:97-98.

Noonan, M. Recent Adaptions of the Devanagari Script for the Tibeto-Burman Languages of Nepal. [Online]. Available: http://crossasia-repository.ub.uni-heidelberg.de/202/ (accessed October 2016).

Oppitz, M. 1973. Myths and Facts: Reconstructing some Data concerning the Clan History of the Sherpa. [Online]. Available: http://himalaya.socanth.cam.ac.uk/collections/journals/kailash/pdf/kailash_02_0102_04.pdf (accessed October 2016)

Prinsloo, D.J. 2011. A Critical Analysis of the Lemmatisation of Nouns and Verbs in isiZulu. In: *Lexikos* 21:169-193.

Prinsloo, D.J. 2015. Corpus-based Lexicography for Lesser-resourced Languages: Maximizing the Limited Corpus. In: *Lexikos* 25:285-300.

Rasmussen, A.G. 1998. Teortisk behandling af indholdet I faglige forklaringer. In: Tarp, S (Ed.). *Leksikografi som speciale. Bind I.* Aarhus: Spansk Institut and Center for Leksikografi, Aarhus School of Business. 127-153. Quoted in: Tarp 2008:79.

Roget, P.M. 1958. Roget's Thesaurus. Harmondsworth, Middlesex: Penguin Books. Quoted in Moe (2001).

Saeed, J.I. 1997. *Semantics*. Second edition. Blackwell Publishing.

Scerba, L.V. 1940. Towards a General Theory of Lexicography. In: *International Journal of Lexicography*. Vol. VIII. Number 4, 1995. Oxford: Oxford University Press. 315-350. Quoted in Tarp (2008):18-20.

Schöttelndreyer, B. 1970. A Devanagari spelling system for the Sherpa Language. Kathmandu: Summer Institute of Linguistics and Tribhuvan University. (Unpublished manuscript)

Schöttelndreyer, H. 1971. A Guide to Sherpa Tone. Summer Institute of Linguistics and Tribhuvan University, Kathmandu. (Unpublished manuscript)

Swanepoel, P. Dictionary typologies: A pragmatic approach. In: Van Sterkenburg, P (Ed.). 2003. *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins Publishing Company:44-69.

Tarp, S. 1998. Leksikografien på egne ben. Fordelingsstrukturer og byggedele i et brugerorienterete perspektiv. *Hemes, Journal of Linguistics No. 21*. 121-137. Quoted in: Tarp 2008:81.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer.

Thomason, S.G. 2015. *Endangered Languages: An introduction*. Cambridge Textbooks in Linguistics. Cambridge: University Press.

Tournadre, N., L.N. Sherpa, G. Chodrak. and G. Oisel. 2009. *Sherpa-English and English-Sherpa Dictionary with Literary Tibetan and Nepali equivalents.* Kathmandu: Vijra Publications.

Tsunoda, T. 2005. *Language Endangerment and Language Revitalization.* Trends and Linguistics Studies and Monographs 148, Berlin: De Gruyter.

UNESCO Ad Hoc Expert Group on Endangered Languages, Language Vitality and Endangerment. 2003. Document submitted to the International Expert Meeting in UNESCO Programme Safeguarding of Endangered Languages. Paris, 10-12 March 2003

Watters, S.A. 1999. Tonal Contrasts in Sherpa. In: Yadava and Glover (Eds.). *Topics in Nepalese Linguistics.* Kathmandu: Royal Nepal Academy:54-77.

Wiegand, H.E. 1983. Was ist eigentlich ein Lemma? In: Wiegand, H.E (Ed.). Studien zur neuhochdeutscen Lexikographie. Hildesheim: Geor Olms Verlag: 401-474. Quoted in: Gouws and Prinsloo 2005:5.

Wiegand, H.E. 1984. On the structure and contents of a general theory of lexicography. In: Hartmann, R.R.K (Ed.). 1983. *LEXeter '83.* Tübingen: Max Niemeyer:13-30.

Wiegand, H.E. 1987. Zur handlungstheoretischen Grundlegung der Wörterbenutzungsforchung. *Lexicographica. International Annual for Lexicography 3,* 178-227. Quoted in: Tarp 2008:29.

Wiegand, H.E. 1989. Der Begriff der Mikrostuktur: Geschichte, Probleme, Perspektiven. In: Hausman, FJ. et al. (Eds.). 1989-1991:409-462. Quoted in: Gouws and Prinsloo (2005:116).

Wiegand, H.E. 1996. A theory of lexicographic texts. An overview. In: SA Journal of Linguistics 14/1:134-149. Quoted in: Gouws and Pinsloo 2005: 57.

Wiegand, H.E. 1996a. Über die Medisostrukturen bei gedruckten Wörterbüchern. In: Zettersten, A. and V.H. Pedersen (Eds.). 1996. *Symposium on Lexicography VIII*. Tübingen: Max Niemeyer: 11-43. Quoted in: Gouws and Pinsloo 2005:177.

Wiegand, H.E. 1998. Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin/New York: de Gruyter. Quoted in: Bergenholtz, Henning and Tarp, Sven:179-181.

Wiegand, H.E. 2002. Translational Equivalence in the Bilingual Dictionary. In: *Lexikos* 12:147-154.

Wiegand, H.E., I. Feinauer. and R.H. Gouws. 2013. Types of Dictionary Articles in Printed Dictionaries. Gouws, Rufus H., Ulrich Heid, Wolfgang Schweickard and Herbert Ernst Wiegand (Eds.). 2013:314-366. Quoted in: Gouws 2014:158.

Weinberg, M. 2009. Sherpa Language Program Summary (A personal report to The Mountain Institute that funded this project).

Zgusta, L. 1971. *Manual of Lexicography*. Den Haag: Mouton.

Zgusta, L. 1989. The role of dictionaries in the genesis and development of the standard. In: Hausmann, F. J. et al. (Eds.). 70-79.

Zgusta, L. 2006. *Lexicography Then and Now*. Selected Essays edited by F.S.F. Dolezal. and T.B.I. Creamer. Tübingen: Max Niemeyer Verlag.

Appendices

1. 240 Word lists for Lexical similarity comparison

1. body
2. head
3. hair
4. face
5. eye
6. ear
7. nose
8. mouth
9. tooth
10. tongue
11. breast
12. belly
13. arm
14. elbow
15. palm
16. finger
17. fingernail
18. leg
19. skin
20. bone
21. heart
22. blood
23. urine
24. feces
25. village
26. house
27. roof
28. door
29. firewood
30. broom
31. mortar
32. pestle
33. hammer
34. knife
35. axe
36. rope
37. thread
38. needle
39. cloth
40. ring
41. sun
42. moon
43. sky
44. star
45. rain
46. water
47. river
48. cloud
49. lightning
50. rainbow
51. wind
52. stone
53. path
54. sand
55. fire
56. smoke
57. ash
58. mud
59. dust
60. gold
61. tree
62. leaf
63. root
64. thorn
65. flower
66. fruit
67. mango
68. banana
69. wheat
70. millet
71. rice
72. potato
73. eggplant
74. groundnut
75. chili
76. turmeric
77. garlic
78. onion
79. cauliflower
80. tomato
81. cabbage
82. oil
83. salt
84. meat
85. fat
86. fish
87. chicken
88. egg
89. cow
90. buffalo
91. milk
92. horn
93. tail
94. goat
95. dog
96. snake
97. monkey
98. mosquito
99. ant
100. spider
101. name
102. man
103. woman
104. child
105. father
106. mother
107. older brother
108. younger brother
109. older sister
110. younger sister
111. son
112. daughter
113. husband
114. wife
115. boy
116. girl
117. day
118. night
119. morning
120. noon
121. evening
122. yesterday
123. today
124. tomorrow
125. week
126. month
127. year
128. old
129. new
130. good
131. bad
132. wet
133. dry
134. long
135. short
136. hot
137. cold
138. right
139. left
140. near
141. far
142. big
143. small
144. heavy
145. light
146. above
147. below
148. white
149. black
150. red
151. one
152. two
153. three
154. four
155. five
156. six
157. seven
158. eight
159. nine
160. ten
161. eleven
162. twelve
163. twenty
164. one hundred
165. who?
166. what?
167. where?
168. when?
169. how many?
170. what kind?
171. this
172. that
173. these
174. those
175. same
176. different
177. whole
178. broken
179. few
180. many

181. all
182. eat, he ate
183. bite, he bit
184. he is,
he was hungry
185. drink, he drank
186. he is,
he was thirsty
187. sleep, he slept
188. lie down,
he lay down
189. sit, he sat
190. give, he gave
191. it burns,
it burned
192. don't die, he
died
193. don't kill,
he killed
194. fly, it flew
195. walk, he walked
196. run, he ran
197. go, he went
198. come, he came
199. speak, he spoke
200. listen, he heard
201. look, he saw
202. I (1st sg.)
203. you
(2nd sg. informal)
204. you
(2nd sg. formal)
205. he
(3rd sg .masculine)
206. she
(3rd sg. feminine)
207. we
(1st pl. inclusive)
208. we
(1st pl. exclusive)
209. you (2nd pl)
210. they (3rd pl)
211. not
212. person
213. bird
214. louse
215. seed
216. bark
217. feather
218. knee
219. neck
220. liver
221. he knew
222. he swam
223. stand, he stood
224. earth
225. snow mountain
226. green
227. yellow
228. full
229. round
230. turquoise
231. yak (male)
232. yak (female)
233. carpet (sitting)
234. fly (n)
235. barley flour
236. horse
237. yak butter
238. butter tea
239. hat
240. silver

2. Expanded Graded Intergenerational Disruption Scale (Simons & Lewis 2010)

(Source: https://www.ethnologue.com/about/language-status)

Level	Label	Description
0	International	The language is widely used between nations in trade, knowledge exchange, and international policy.
1	National	The language is used in education, work, mass media, and government at the national level.
2	Provincial	The language is used in education, work, mass media, and government within major administrative subdivisions of a nation.
3	Wider Communication	The language is used in work and mass media without official status to transcend language differences across a region.
4	Educational	The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.
5	Educational	The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.
6a	Vigorous	The language is used for face-to-face communication by all generations and the situation is sustainable.
6b	Threatened	The language is used for face-to-face communication within all generations, but it is losing users.
7	Shifting	The child-bearing generation can use the language among themselves, but it is not being transmitted to children.
8a	Moribund	The only remaining active users of the language are members of the grandparent generation and older.
8b	Nearly Extinct	The only remaining users of the language are members of the grandparent generation or older who have little opportunity to use the language.
9	Dormant	The language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency.
10	Extinct	The language is no longer used and no one retains a sense of ethnic identity associated with the language.